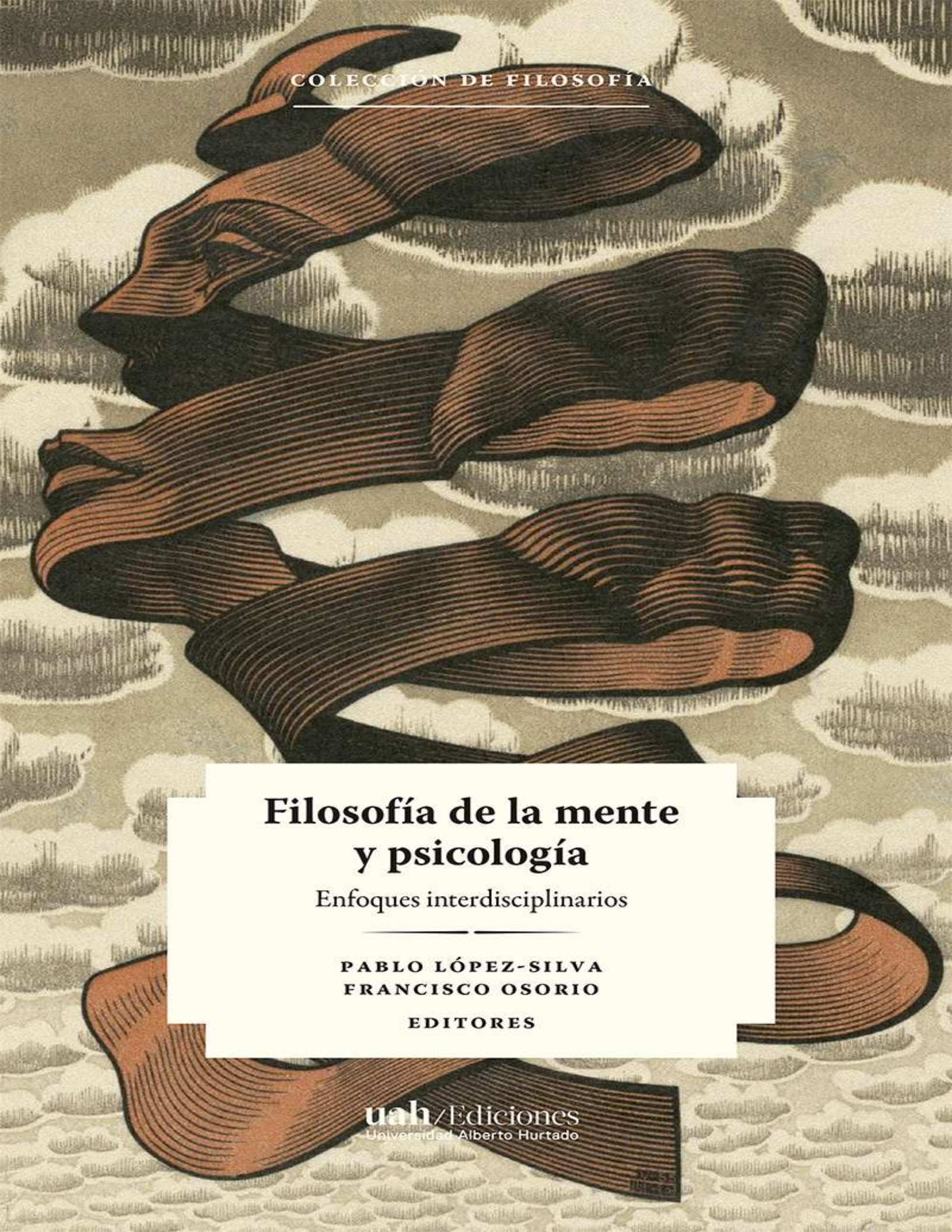


COLECCIÓN DE FILOSOFÍA

Filosofía de la mente y psicología

Enfoques interdisciplinarios

PABLO LÓPEZ-SILVA
FRANCISCO OSORIO
EDITORES

uah/Ediciones
Universidad Alberto Hurtado

Filosofía de la mente y psicología

Enfoques interdisciplinarios

Filosofía de la mente y psicología
Enfoques interdisciplinarios

Pablo López-Silva
Francisco Osorio
Editores

Ediciones Universidad Alberto Hurtado
Alameda 1869 – Santiago de Chile
mgarciam@uahurtado.cl – 56-228897726
www.uahurtado.cl

Marzo de 2020

ISBN libro impreso: 978-956-357-235-3
ISBN libro digital: 978-956-357-236-0

Este texto fue sometido al sistema de referato ciego.

Coordinador Colección Filosofía
Francisco Pereira Gandarillas

Dirección editorial
Alejandra Stevenson Valdés

Editora ejecutiva
Beatriz García-Huidobro

Diseño interior y portada
Francisca Toral R.

Imagen de portada: Corteza, de M. C. Escher

Con las debidas licencias. Todos los derechos reservados. Bajo las sanciones establecidas en las

leyes, queda rigurosamente prohibida, sin autorización escrita de los titulares del copyright, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, así como la distribución de ejemplares mediante alquiler o préstamos públicos.

Diagramación digital: ebooks Patagonia
www.ebookspatagonia.com
info@ebookspatagonia.com

Filosofía de la mente y psicología

Enfoques interdisciplinarios

PABLO LÓPEZ-SILVA
FRANCISCO OSORIO
EDITORES

uah/Ediciones
Universidad Alberto Hurtado

Índice

Prólogo

El carácter elusivo del concepto de mente

Tim Bayne

Presentación

La continuación de la filosofía de la mente en el siglo XXI

Pablo López-Silva y Francisco Osorio

Las grandes preguntas

Capítulo I

¿Qué es la mente?

Alfredo Gaete

Capítulo II

La conciencia en la discusión actual: cuestiones filosóficas

Marcelo Díaz

Capítulo III

Auto-referencia y el problema de la conciencia en la vida y obra de Francisco Varela

Roberto Ferrari y Ricardo Pulido

Problemas específicos en filosofía de la mente

Capítulo IV

La crítica de Hans Jonas al monismo materialista

Alejandro Serani

Capítulo V

Género, imitación e inteligencia: una revisión crítica del enfoque funcionalista de Alan Turing

Rodrigo González

Capítulo VI

Las intenciones como compromisos: un análisis de las posturas de la doble perspectiva sobre la agencia humana

María Ayelén Sánchez

Enfoques para el estudio de lo mental

Capítulo VII

Ciencia cognitiva como multidisciplina

Francisco Osorio

Capítulo VIII

Elementos para una aproximación científica al estudio de la conversación

Edmundo Kronmüller

Cuando la mente se altera

Capítulo IX

La aproximación fenomenológica en psiquiatría

Otto Dörr

Capítulo X

Esquizofrenia y la naturaleza de lo mental: sobre el estudio interdisciplinario de las patologías mentales

Pablo López-Silva

Capítulo XI

El problema de la explicación de los delirios: en defensa de un enfoque multifactorial

Emilia Vilatta

Autoras y autores

PRÓLOGO

El carácter elusivo del concepto de mente

En 1996 y 1997 una serie de memorables partidas de ajedrez se llevaron a cabo en los Estados Unidos. En dos rondas de seis juegos cada una, el maestro Gary Kasparov se enfrentó a la supercomputadora Deep Blue. Como es sabido, Kasparov ganó la primera ronda mientras que Deep Blue se quedó con la segunda. Además de generar un sinnúmero de preguntas sobre las distintas estrategias desplegadas por ambos participantes, el evento también motivó diversas preguntas de carácter filosófico.

Claramente, existían muchas diferencias físicas y funcionales entre Kasparov y Deep Blue. Sin embargo, en este contexto cabe preguntarse ¿en qué grado tales diferencias estuvieron acompañadas de diferencias psicológicas significativas? Ambos contrincantes parecían seguir las reglas estándares del ajedrez y la manera magistral en que ambos jugaron parecía sugerir un alto nivel de inteligencia. Viendo esto, uno podría fácilmente sentirse tentado a pensar que Deep Blue poseía el mismo tipo de atributos psicológicos que Kasparov. Sin embargo, tal aseveración sería tomada por muchos como implausible e incluso como hasta contraintuitiva. Claramente, Kasparov parecía entender lo que estaba haciendo, pero ¿podríamos decir lo mismo en el caso de Deep Blue? Indudablemente, Kasparov evaluó los méritos de todas las posibles movidas disponibles para atacar a su oponente, pero, ¿hizo Deep Blue lo mismo? Finalmente, Kasparov estaba consciente de lo que estaba haciendo, pero ¿deberíamos decir lo mismo en el caso de Deep Blue?

Los conceptos que comúnmente asociamos al término “mente” (inteligencia, entendimiento, consciencia, pensamiento, entre muchos otros) son sumamente enigmáticos. Así, a pesar de que los seres humanos parecemos disfrutar de una especie de familiaridad primitiva con todos estos, al mismo tiempo, tales conceptos son elusivos en diversas maneras. El carácter elusivo del concepto de mente se hace más claro en el contexto de la comprensión de los diversos síntomas que caracterizan sus patologías

(delirios, fobias y las alucinaciones, por ejemplo). ¿Cómo podemos entender tales estados psicopatológicos? ¿Cómo podría el estudio de estos estados iluminar el entendimiento general de los estados mentales no patológicos con los cuales ya estamos familiarizados en nuestro diario vivir?

Pues bien, a pesar de que el desarrollo práctico y teórico en los campos asociados al estudio de la mente ya posee una privilegiada historia en Latinoamérica, este ha recibido un renovado ímpetu durante los últimos años con la emergencia de una nueva generación de investigadores. Los miembros de esta generación no se confinan al trabajo aislado en sus propias disciplinas, sino que circulan libremente entre psicología, filosofía, antropología y ciencias experimentales. Ninguna colección puede expresar todo el progreso realizado en el estudio de la mente en los últimos años en Latinoamérica; sin embargo, nuestra esperanza es que este libro logre capturar tal espíritu. El principal objetivo de la edición de esta colección debe ser entendido no solamente como una introducción a algunas de las discusiones más importantes en el estudio de la mente humana, sino también como una forma de motivar a que las futuras generaciones de investigadores en Latinoamérica se atrevan a desarrollar sus propias contribuciones.

Tim Bayne

PRESENTACIÓN

La continuación de la filosofía de la mente en el siglo XXI

Pablo López-Silva y Francisco Osorio

Este libro es como una carrera de relevos, donde el testimonio entregado pasa de una generación de académicos a la siguiente para continuar con el desarrollo de la investigación respecto de los fenómenos mentales y sus problemas asociados. En esta época de transición en que nos encontramos, la ciencia del siglo XX entrega sus saberes y prácticas a los académicos del siglo XXI, coexistiendo ambas generaciones durante este lapso de tiempo. El cambio de siglo, así como el cambio de generación de investigadores, no hace sino mostrar en su dinámica lo que es estable desde los inicios de la filosofía, esto es, la pregunta sobre aquello que llamamos "mente" y todos sus derivados. Como una estrella que puede ser vista de ambos hemisferios, el fenómeno de la conciencia, la inteligencia, la agencia, las patologías mentales y la pregunta misma de cómo investigar todo lo anterior, se presentan como una luz que nos atrae de tiempos inmemoriales. Por lo tanto, estos escritos continúan una antigua conversación, incorporando a los próximos participantes en esta mesa académica.

El estudio de la mente humana debe ser una de las empresas más desafiantes en la actualidad. Con la publicación de *Mind y Cosmos*, Thomas Nagel nos remece al sugerir que aún estamos muy lejos de entender la forma en que nuestra mente surge en un universo material. Esto no solo debe ser visto como un problema, sino como una verdadera oportunidad para profundizar en esta discusión. Ahora bien, lejos están aquellos años donde los filósofos construían sus teorías en sus castillos de cristal. A la luz de las revoluciones cognitivas y los avances en neurociencias, los filósofos cada vez han ido explorando más disciplinas con el fin de alimentar sus ideas respecto de lo mental. De esta forma, el estudio de la mente humana

se constituye necesariamente como una empresa interdisciplinaria; un campo de debate que se enriquece no solo desde el razonamiento lógico, sino que también desde la evidencia empírica en psicología, psiquiatría, biología y las ciencias sociales. Sin duda esta interdisciplinariedad no solo conlleva a complejizar los enfoques en filosofía y ciencias de la mente, sino que también a la proliferación de diversas preguntas que nacen, exactamente, desde tal intersección disciplinar.

La presente compilación es testigo de ambas dimensiones en el avance de la filosofía de la mente en la actualidad. Por una parte, los capítulos que contiene son una expresión de la complejidad que nace al incluir saberes de diversas disciplinas al discutir un tema específico. Por otra parte, la diversidad de preguntas tratadas constituye una continuidad entre aquellas discusiones clásicas y aquellos debates que surgen de la multidisciplinariedad propia de la filosofía de la mente.

Si tuviésemos que agrupar consistentemente las contribuciones de esta compilación en secciones, la primera podría ser denominada “Las grandes preguntas”. Esta sección inicia con la discusión sobre el mismo problema asociado al concepto de mente. No es arriesgado indicar que el problema mente-cuerpo o el cómo logramos formular la relación entre lo mental y material es uno de los problemas madres de la filosofía de la mente. Con el fin de introducirnos a este asunto, Alfredo Gaete explora diversos usos que tiene el concepto de mente en la literatura dominante en psicología y filosofía. Surge acá la pregunta respecto de si realmente podemos atribuir una característica común a todos los usos del concepto en la literatura. Aquellos que defiendan una respuesta positiva enfrentan un proyecto sumamente complejo, esto es, el de unificar dimensiones agenciales, subjetivas, funcionales e intencionales del concepto. Aquellos que defienden una respuesta negativa tampoco tienen un camino fácil; estos enfrentan a la necesidad de defender su posición de las consecuencias que podría tener para el mismo proyecto de estudiar lo mental y la unificación de la ciencia. ¿Qué será lo que nos quiere decir el autor al final de su contribución? En el segundo y tercer capítulo de esta sección, Marcelo Díaz, Roberto Ferrari y Ricardo Pulido exploran otra de las grandes preguntas en la filosofía de la mente, la conciencia. Imitando la estrategia de Gaete, Marcelo Díaz ofrece una completa identificación y análisis de los usos del término en la literatura en psicología y filosofía de la mente. Ahora

bien, este análisis servirá para iluminar –a la luz de las ideas de Wittgenstein– el asunto sobre cómo ciertos hábitos lingüísticos podría generar ilusiones conceptuales que, a su vez, podrían constituir la base de ciertos problemas filosóficos. Ahora bien, Ferrari y Pulido proponen otro enfoque para abordar la conciencia. Los autores ocuparán la obra de Francisco Varela, en especial el concepto de auto-referencia, como un filtro para debatir algunas de las dimensiones más importantes del fenómeno de la conciencia. La pregunta de Varela sobre el estatuto de la conciencia en el universo es, simplemente, de gran interés para nosotros, pues nos arroja una mirada diferente en la comunicación interdisciplinaria sobre la mente o la conciencia y logra conectar preguntas clásicas de la filosofía, como *insights* provenientes de la historia local de las ideas filosóficas en Latinoamérica. Finalmente, esta sección concluye con el análisis que Alejandro Serani ofrece sobre el monismo materialista, esto, como una instancia específica de una teoría ontológica de aquellos que llamamos mente. La idea de este capítulo es no solo ofrecer debates respecto de los conceptos, sino que también de los marcos filosóficos que intentan darle sentido a las “grandes preguntas” en filosofía de la mente.

La segunda sección de esta compilación se focaliza en “Problemas específicos en filosofía de la mente”, esto es, en el debate de conceptos o fenómenos que surgen del funcionamiento de aquello que llamamos mente. En este contexto, Rodrigo González se adentra en la discusión contemporánea sobre la posibilidad de la inteligencia de máquina. Uno de los tantos usos que posee el concepto de mente se asocia con la inteligencia, así, surgen preguntas respecto de si organismos con diferente biología (animales no-humanos) o materialidad (robots) podrían desarrollar el tipo de inteligencia que nosotros gozamos. Para muchos, esta pregunta es problemática, por lo que se ha re-formulado indicando que el verdadero desafío para los humanos sería definir el tipo de acciones que una máquina –por ejemplo– debería realizar para hacernos creer que es inteligente. Una de las fortalezas de la contribución de González está en la introducción de una variable comúnmente ignorada en el debate, esto es, el género. Así, el autor critica la idea de cómo la imitación de inteligencia de la forma hacerse-pasar-por es insuficiente para la acción inteligente desde el punto de vista de la ontología social. El debate sobre la inteligencia de máquina sin duda invita a reflexionar sobre asuntos como la voluntad y la agencia

humana; es más, algunas personas indican que aquello que nos distingue de las máquinas es la realización de actos voluntarios. Tomando estas preguntas en cuenta, la contribución de Ayelén Sánchez analiza diversas teorías que intentan darle sentido al concepto de agencia humana. En su primera instancia, el proyecto de Sánchez es clarificatorio, al intentar formular de forma clara el enfoque que conecta causalmente la acción con la intención. En su segunda parte, el proyecto deviene revisionista al indicar que este enfoque posee varios problemas para realmente lidiar con el problema ontológico de la agencia.

Una vez desarrollada la sección que intenta lidiar con debates específicos, la compilación ofrece dos contribuciones que podrían asociarse en torno a la idea de "Enfoques para el estudio de lo mental". Francisco Osorio, desarrolla una pregunta fundamental para efectos de la compilación misma. En su capítulo "Ciencia cognitiva como multidisciplina", el autor revisa la pregunta sobre qué significa que lo mental sea estudiado multidisciplinariamente, lo que, en consecuencia, llevará al autor a discutir el asunto sobre qué y cómo sería una ciencia cognitiva que se autodenominada como multidisciplinaria. Este capítulo surgirá de la invitación de varios autores de esta compilación a pensar un trabajo de investigación desde diferentes miradas, pero frente a lo cual hay que evaluar modelos de cómo proceder, pues debemos trabajar en la práctica, de alguna manera, con diferentes teorías, métodos y disciplinas de origen. Por su parte, el capítulo ofrecido por Edmundo Kronmüller se centra en un fenómeno fundamental para la existencia humana, pero que sin duda ha sido pasado por alto por los grandes proyectos investigativos del siglo XX. El autor intentará esbozar algunos de los aspectos más importantes para el estudio científico de la conversación humana, un trabajo que sin duda aportará dimensiones relevantes para el desarrollo de esta línea de estudio a niveles locales y comunes. Así, es interesante ver cómo para el autor cada conversación es única en sí misma, pero al mismo tiempo es todas las demás conversaciones a la vez. Esta idea, muy interesante, no es sino una aproximación científica a los fenómenos que ocurren en una conversación. Ahí se encuentran la incertidumbre, los elementos comunes entre dos personas que tratan de comunicarse, las referencias (que no siempre son tan claras) y todo el proceso cognitivo que ello implica. Investigar la conversación es investigar la inmediatez, la certeza, la duda, la interacción

humana en su máxima expresión, y a la vez, es estudiar uno de los aspectos evolutivos más importantes para la adaptación de nuestra especie.

La sección final del libro se enfoca en aquellas discusiones que surgen cuando la mente se altera. Es sumamente importante entender que un estudio interdisciplinario de lo mental no puede dejar afuera el estudio de lo patológico como forma de entender lo normal. De esta forma, esta sección explora algunos aspectos de los transtornos psiquiátricos para ampliar el alcance de nuestra colección. Primero Otto Dörr nos pone en el contexto del siglo XX y desde ahí, específicamente de la mano de la fenomenología y la hermenéutica, nos presenta los principales problemas de estudio de la psiquiatría, junto a los principales académicos que trazaron los contornos de la investigación de su siglo. El fin de esta contribución es ofrecer una contextualización respecto de cómo la fenomenología puede nutrir el estudio de lo mental en sus dimensiones normales y patológicas. La siguiente contribución, escrita por Pablo López-Silva, continúa en la senda de la pregunta sobre lo patológico en el contexto de lo mental. En su caso, con un enfoque específico en el estudio de la esquizofrenia. El autor utiliza esta patología para explorar las diferencias entre lo mental y lo físico y, sobre todo, para discutir algunos aspectos fundamentales de la relación entre conciencia y autoconciencia en el contexto de las patologías mentales. La propuesta de López-Silva será que debemos considerar un enfoque interdisciplinario para abordar estas preguntas. La contribución final de la compilación se constituye como una profundización de la última sección y como una ejemplificación de cómo la investigación interdisciplinaria en filosofía y ciencias de la mente podría realizarse. Así, Emilia Vilatta se centra en la forma en que debiésemos entender la psicogénesis de los delirios, uno de los síntomas claves de la psicosis. Mediante una evaluación sistemática de las teorías dominantes, el capítulo sugiere que estas solo presentan visiones incompletas y que el mejor prospecto para la comprensión del fenómeno emerge desde un entendimiento multifactorial. Sin duda, esta idea deja un desafío abierto, el desafío de profundizar la investigación de lo mental desde un punto de vista necesariamente interdisciplinario.

Quisiéramos ahora finalizar con algunas palabras respecto de la compilación en sí. La motivación inicial para embarcarnos en esta empresa surge el año 2013 en el contexto de un seminario organizado por la revista

Cinta de Moebio llamado “Enfoques Interdisciplinarios en Filosofía de la Mente y Psicología”. La construcción de ese espacio dejó en claro la necesidad de crear conexiones más fuertes y perdurables, asociadas a productos concretos, y de participación continua en torno a la investigación en filosofía y ciencias de la mente. A su vez, esta instancia visualizó la diversidad de ideas originales de alto nivel en el medio local. Desde ese momento, se comenzó a buscar autores interesados en la construcción de un espacio común que pudiese expresar el espíritu innovativo e interdisciplinario que promovió ese seminario. Al pasar de los años, tales espacios fueron complejizándose y el interés comenzó a crecer, lo que finalmente nos lleva al presente producto final, una compilación interdisciplinaria de trabajos que tratan diversas temáticas conceptuales en torno a una comprensión cooperativa de la mente humana. Con justicia, el mayor mérito en la producción final de la compilación es para Pablo López-Silva, quien se llevó buena parte del trabajo. A través de esta breve presentación damos las gracias a todos quienes creyeron en este proyecto y a sus lectores les damos un profundo agradecimiento por su interés en estos modestos
pensamientos.

LAS GRANDES PREGUNTAS

CAPÍTULO I

¿Qué es la mente?

Alfredo Gaete

Introducción

A menudo no está claro qué quieren decir los filósofos y los psicólogos cuando hablan de la “mente” y “lo mental”, a pesar de que hablan mucho de ello. Mientras algunos se sienten conceptualmente forzados a oponer lo mental a lo físico, otros no; mientras algunos asocian la mente principalmente al intelecto, otros la ven más relacionada con la sensación y el sentimiento, o con la acción, o con todo eso por igual, o con otras cosas. De modo que cuando se dice, por ejemplo, que las creencias y los dolores son estados mentales, o que las imaginaciones ocurren en la mente de la gente, o que tal persona está mentalmente enferma, etc., no es del todo evidente qué es exactamente lo que se está señalando. Ni siquiera es evidente que en todos esos casos se esté expresando un mismo concepto: por ejemplo, que el concepto freudiano de mente sea el mismo que el de Piaget; o que quienes afirman (con o sin razón) que algunas máquinas pueden ejecutar operaciones mentales, entiendan ese adjetivo en el mismo sentido en que se dice que la ansiedad y otras sensaciones son mentales.

En este trabajo exploro la posibilidad de que no exista unidad conceptual entre algunos de los diversos usos de “mente” y “mental”; o, lo que es lo mismo, la posibilidad de que tales palabras tengan más de un significado. Si así fuera, decir que los dolores y las creencias son estados mentales sería tan equívoco o ambiguo como decir que los asientos de la plaza y las instituciones financieras son bancos, con la diferencia de que, en el primer caso, la ambigüedad no sería para nada obvia. Trataré también de explicar

esta falta de obviedad.

Sobre los usos de “mente” y “mental”

Primero, identifiquemos los principales usos que tienen los términos “mente” y “mental” en la psicología y la filosofía de la mente contemporánea. Me voy a referir a cuatro solamente, pero hay más –y ha habido otros en el pasado–[1]. Es necesario indicar que acá deseo hablar de “usos” y no de “sentidos” para no dar por supuesto de antemano aquello que quiero argumentar, esto es, que algunos de estos usos corresponden no solo a inflexiones de un mismo concepto sino a sentidos o significados totalmente distintos de los términos.

Partamos por lo que podríamos llamar el “uso subjetivo” de las palabras “mente” y “mental”. En este uso, el sello de lo mental es la “conciencia fenoménica”. Con esto me refiero a esa propiedad que tienen todas las experiencias conscientes y solo ellas –esa fenomenología subjetiva distintiva que está presente en todo experienciar–[2]. Esta propiedad es, a juicio de muchos filósofos, aquello que hace mental a lo mental. Strawson (2004), por ejemplo, ha señalado recientemente que “los fenómenos experienciales son los únicos fenómenos verdaderamente mentales” (p. 295). Por cierto que sí, en tanto nos mantengamos en el uso subjetivo del término. Es este uso lo que hace inteligible el hecho de que Putnam (1980) haya empezado su ya clásico ensayo “La naturaleza de los estados mentales” con la afirmación de que las preocupaciones típicas del filósofo de la mente pueden expresarse en tres preguntas sobre el dolor, a saber: “(1) ¿Cómo sabemos que otra gente tiene dolores? (2) ¿Son los dolores estados cerebrales? (3) ¿Cuál es el análisis del concepto de dolor?” (Putnam 1980, p. 223). Y es contra este mismo uso, o contra la pretensión de ubicarlo como el único admisible, que Freud tuvo que combatir en más de un trabajo para posicionar su teoría psicológica en el mundo académico, que estaba altamente permeado por una resistencia filosófica (principalmente de corte cartesiano) a la posibilidad de una mente inconsciente. Como sea, la idea es que, en muchos contextos, poseer mente significa tener la capacidad de tener experiencias conscientes, es decir, una vida subjetiva –y, por ende, un punto de vista y un acceso “directo” a ciertos contenidos (por nombrar

algunas de las particularidades que vienen incluidas en la subjetividad)–.

Otro uso de "mente", considerablemente distinto de recién revisado, identifica la mente con la inteligencia. Según algunos, este uso se relaciona estrechamente con el concepto de mente más cotidiano, más de sentido común (Feigl, 1967; Ryle, 1963) –como cuando se dice de alguien muy inteligente que tiene una "mente brillante"–[3]. Pero también refleja, creo, un uso bien difundido en la psicología. Por ejemplo, cuando Piaget habla de la mente del niño parece referirse al pensamiento del niño, a sus habilidades cognitivas, a su capacidad de modificar el medio y de automodificarse, etc. (y no, por ejemplo, a su consciencia fenoménica). Asimismo, para el psicólogo cognitivo de la segunda mitad del siglo XX, la mente no tiene que ver tanto con los dolores o las picazones u otros estados experienciales como con ciertas capacidades que, como señalan algunos partidarios de la teoría computacional de la mente, podrían eventualmente llegar a darse en una máquina u otro tipo de ser no sintiente. Esto es polémico, por cierto; pero el punto relevante aquí es que buena parte de nuestro discurso –tanto cotidiano como científico– sobre la mente y lo mental tiene que ver con la inteligencia.

Permítaseme hablar aquí del "uso agencial" de "mente" y "mental", en el entendido de que la inteligencia es, ante todo, un fenómeno que tiene que ver con la capacidad de un agente de conducir su comportamiento de manera racional, esto es, motivado por razones o fines[4]. En claro contraste con el uso subjetivo, en este caso atribuirle mente a un ser no es decir que "le pasan" ciertas cosas, sino que "hace" o "sabe hacer" ciertas cosas. Es atribuirle un cierto tipo de agencia: la agencia racional o inteligente. De modo que al caracterizar a las personas como dotadas de mente en este uso del término, lo que queremos decir es que las personas son agentes inteligentes, agentes que saben o entienden cómo guiar su conducta, que actúan de acuerdo a razones, que entran en el terreno de la acción voluntaria y la racionalidad práctica, etc. Y es esta dimensión racional-práctica de nuestra existencia la que, como he dicho, muchos psicólogos han querido describir y explicar en sus teorías sobre la mente humana.

Un tercer uso de mente podría describirse como el "uso intencional" o "representacional" (véase, por ejemplo, Crane, 1998; 2001; Davidson, 2001; Kim, 1971). Aquí lo mental es aquello que se refiere o está dirigido a otra cosa, aquello que tiene significado, o contenido, o –más técnicamente–

“objeto intencional”. La historia se remonta a Brentano (1995) y su sugerencia de que todo fenómeno mental es intencional en este sentido; de modo que, por ejemplo, en el juicio algo se afirma o se niega, en el amor algo se ama, en el odio algo se odia, en el deseo algo se desea, y así sucesivamente. Según Brentano (1995), la intencionalidad así entendida es “característica exclusiva de los fenómenos mentales”, de modo que “podemos [...] definirlos fenómenos mentales diciendo que son aquellos fenómenos que contienen intencionalmente un objeto dentro de ellos”.

Se dice que Freud entendía lo mental de esta manera (ver por ejemplo, Flanagan 1991). La sugerencia es plausible, considerando la importancia del mundo del significado en la teoría freudiana. Pero lo mismo nos podría acercar al uso agencial. Quizá ambos usos expresen, después de todo, un mismo concepto. Si me apuran, yo diría que el acceso al mundo del significado es exclusivo de los seres inteligentes y, en consecuencia, la intencionalidad presupone inteligencia. Como sea, tanto el uso agencial como el intencional dejan abierta la posibilidad de que haya estados mentales inconscientes, que es lo primero que necesitaba Freud para que su teoría psicológica no resultara auto-contradictoria desde el principio.

Consideremos todavía un cuarto uso de “mente” y “mental” que se ha extendido bastante en la escena psicológica y filosófica desde la segunda mitad del siglo pasado, sobre todo después de la revolución cognitiva. Me refiero al “uso funcionalista”, según el cual lo mental está dado por ciertos procesos y estados que se encuentran a mitad de camino en la cadena causal que va desde el estímulo a la respuesta. Una mujer muy guapa aborda al señor Pérez y le pregunta si le gustaría ir al cine con ella. Todo lo que ocurre “dentro” del señor Pérez como consecuencia de esto y que a su vez tiene como consecuencia la respuesta del señor Pérez (digamos que se sonroja y tartamudea algo), todo eso califica como mental en este uso. Digo que se trata del uso funcionalista porque fueron los funcionalistas, encabezados por pensadores como Jerry Fodor, entre otros, los que popularizaron este uso del término tras la revolución cognitiva que aplastó a la psicología conductista. Eso no quita que los funcionalistas usen “mente” y “mental” para referirse a fenómenos intencionales o subjetivos o que se relacionan con la inteligencia, pero si lo hacen es porque conciben esos fenómenos como jugando un rol causal en la generación de la conducta.

Tenemos, entonces, cuatro usos relativamente vigentes para los términos en cuestión: el subjetivo, el agencial, el intencional y el funcionalista. Mucho más interesante que esta constatación, sin embargo, es la pregunta de si estos cuatro usos son variaciones de una misma idea, de un mismo núcleo conceptual por así decir, o si, como sugerí anteriormente, podría haber más de un concepto de mente.

En favor de la segunda opción puede decirse lo siguiente. Cuando hablamos de "aritmética mental", o de "imágenes mentales", o de "repasar mentalmente" los pasos a seguir para hacer algo, por lo general nos referimos a experiencias subjetivas, a cosas que ocurren en la consciencia de una persona –o sea funcionamos con el uso subjetivo de los términos–. Pero sucede que muchas operaciones que califican como mentales en este uso podrían perfectamente no calificar en otros. El mero hecho de tener imágenes mentales, por ejemplo, no implica ni supone inteligencia. Las imágenes se pueden tener sin razón y combinar estúpidamente. A la inversa, una persona puede ejecutar operaciones inteligentes sin pasar por ningún proceso interno de imaginería o de hablar con ella misma. De hecho, y tal como hizo notar Ryle (1963), la aritmética hecha a lápiz y papel puede ser incluso más inteligente que la aritmética "interna" o "mental" en el uso subjetivo. Lo mental en ese uso tiene poco o nada que ver con lo mental en el uso agencial, salvo quizá por el hecho de que la psicología se ha interesado por ambos tipos de fenómenos. Para Ryle (1963), la distinción entre hablar con sentido y decir tonterías, o entre pensar lo que uno dice y simplemente decir, es transversal a la distinción entre hablar en voz alta y hablar para uno mismo. Esta última distinción tiene que ver con la diferencia entre lo objetivo y lo subjetivo, mientras que la primera tiene que ver con la diferencia entre la agencia inteligente y la meramente refleja, instintiva o automática (Altman, 1997) o, si se prefiere, entre las acciones propiamente tales y la mera actividad (Velleman, 2000).

Puesto que se trata de dos distinciones independientes entre sí, de manera que lo inteligente no presupone ni implica lo subjetivo y, a la inversa, lo subjetivo no presupone ni implica inteligencia, no hay ninguna continuidad lógica o conceptual entre el uso agencial y el uso subjetivo de "mente" y "mental". Dicho de otro modo, se trata de términos ambiguos. Oraciones

como “su mente está perturbada” o “las emociones son estados mentales” pueden ser tan ambiguas como la oración “he visto un banco”; todas ellas pueden expresar (al menos) dos proposiciones totalmente distintas, tales que la obtención de una es independiente de la obtención de la otra. Quien dice, por ejemplo, que la mente de alguien está perturbada, ¿está hablando de la inteligencia de esa persona o de su vida subjetiva? ¿O de ambas cosas? ¿O de ninguna?.

El uso intencional, en cambio, me parece más el resultado de una variante del concepto de inteligencia. Como ya he sugerido, el mundo del significado y la inteligencia podrían estar lógicamente conectados. Si la intencionalidad presupone entendimiento –si, por ejemplo, se excluye que un termómetro a base de mercurio pueda tener representaciones genuinas de la temperatura, dado que no entiende lo que hace o a qué se refiere lo que hace[5]– entonces los conceptos de intencionalidad e inteligencia parecerían estar emparentados. Pero si se acepta que los termómetros representan la temperatura o, por poner otro ejemplo, que los anillos interiores del tronco de un árbol representan o significan la edad del árbol –si se acepta, esto es, que hay contenido intencional en situaciones de esta índole– la conexión lógica o conceptual entre intencionalidad e inteligencia queda borrada de golpe y, en ese caso, habría que agregar un tercer concepto de mente según el cual predicar propiedades mentales no sería equivalente a predicar ni subjetividad ni racionalidad.

Finalmente, el uso funcionalista también parece ofrecer una distinción que no está ligada semánticamente a ninguna de las anteriores. Lo mental, en este uso, es un constructo teórico postulado para explicar cierto tipo de procesos, independientemente de si ese constructo nos reconduce a operaciones inteligentes o a estados intencionales o a experiencias subjetivas o a cualquier otra cosa. Tener mente, para el funcionalista, es tener algo así como un procesador, un dispositivo que media causalmente entre el ambiente y las operaciones de un agente. En ese sentido, tiene que ver con lo interno, pero no necesariamente con lo subjetivo; con la agencia, pero no necesariamente con la inteligencia; y con la sensibilidad al medio, pero no necesariamente con la intencionalidad en el sentido demandante del término. No hay, pues, unidad conceptual entre la mente así concebida y ninguno de los sentidos antes especificados.

En síntesis, si comparamos los usos subjetivo, agencial y funcionalista,

encontramos que las distinciones que subyacen a ellos no parecen derivarse entre sí. Son, quizás, distinciones transversales, independientes, tan independientes como la distinción que haría un mueblista entre un banco y una mesa, o como la que haría un economista entre un banco y un supermercado. La mente de Piaget podría no ser la de los fenomenólogos y quizá ninguna de ellas sea, tampoco, la de Fodor. Pensar que sus teorías deben ser sobre el mismo fenómeno o conjunto de fenómenos ya que todas son "sobre la mente", es como pensar que entre el banco de la plaza y la institución financiera debe haber algún tipo de unidad conceptual.

Conclusión

¿Por qué se ha mantenido oscurecida esta ambigüedad? ¿Por qué no resulta igual de evidente a como resulta la ambigüedad de "banco"? Una posible explicación de esto puede hallarse en el hecho de que la mente humana, en cualquiera de los sentidos antes considerados, involucra de hecho la amplia gama de fenómenos que califican como mentales en alguno (o algunos) de esos usos. El concepto de inteligencia no presupone o implica ni el concepto de subjetividad ni las teorías funcionalistas, pero todos los estados mentales que exhiben inteligencia humana consisten parcialmente en, o son el resultado de, estados subjetivos (sensaciones, percepciones, emociones, etc.). También constituyen parte importante de –si no todo– el conjunto de fenómenos que, a ojos del funcionalista, median causalmente entre estímulo y respuesta en el caso de los seres humanos (que es, por lo demás, el caso que más le interesa explicar). De ahí la dificultad de establecer, por ejemplo, qué concepto o conceptos de mente están en juego en las ciencias cognitivas, y cuándo se apela a uno u otro.

También está el hecho de que la palabra "conciencia" se usa para referir fenómenos bien distintos, no siempre relacionados con la experiencia. Yo tengo conciencia de muchas de mis creencias, sin tener ninguna experiencia de ellas. También tengo conciencia de que tengo conciencia, pero la primera al menos no es fenoménica (en el sentido descrito al principio). Un uso lato de esta palabra podría incluir varios de los sentidos de "mente" antes

expuestos (y otros para los que no hubo espacio también), lo cual a su vez podría llegar a sugerir una unidad conceptual.

Por último, hay un elemento histórico no menos importante que vale la pena mencionar. Cuando Descartes describe las actividades mentales incluye en su lista operaciones cognitivas (p. ej.: dudar, pensar) tanto como estados experienciales (p. ej.: sentir). Por mucho que estos fenómenos mentales sean mentales en distintos sentidos de la palabra, el mero hecho de que filósofos y otros pensadores los vengan agrupando en una misma familia desde hace más de cuatro siglos ha colaborado con la propensión a tratar esta contingencia histórica como una especie de necesidad lógica. Espero haber dado aquí motivos suficientes como para levantar dudas respecto de cuán justificada está esa inclinación.

Referencias

Altman, I. (1997). *The concept of intelligence*. Maryland: University Press of America.

Block, N. (1995). "On a confusion about a function of consciousness". *Behavioral and Brain Sciences 18* (2): 227-88.

Brentano, F. (1995). *Psychology from an empirical standpoint*. London: Routledge.

Crane, T. (1998) "Intentionality as the mark of the mental". En A. O'Hear (ed.) *Contemporary Issues in the Philosophy of Mind.* Cambridge: Cambridge University Press.

—— (2001). *Elements of mind*. Oxford: Oxford University Press.

Davidson, D. (2001). Essays on actions and events. Oxford: Oxford University Press.

Feigl H. (1967). *The "mental" and the "physical". The essay and a postscript.* Minneapolis: University of Minnesota Press.

Flanagan, O. J. (1991). *The science of the mind* (2nd rev. ed.). Cambridge, MA: MIT Press.

James, W. (1950).*The principles of psychology, Vol. 1*. New York: Dover Publications.

Kim, J. (1971). "Materialism and the mark of the mental". *Synthese*22: 323-345.

MacIntyre, A. y Nowell-Smith, P. H. (1960)."Purpose and intelligent action". *Proceedings of the Aristotelian Society, Supp. Vol.* 34: 79-112.

Putnam, H. (1980). "The nature of mental states". En N. Block (ed.), *Readings in philosophy of psychology.* London: Methuen and Co.

Rorty, R. (1979). *Philosophy and the mirror of nature*. New Jersey: Princeton University Press.

Ryle, G. (1963). *The concept of mind*. Harmondsworth: Penguin.

Strawson, G. (2004). "Real intentionality". *Phenomenology and the Cognitive Sciences* 3: 287-313.

Tye, M. (1995).*Ten problems of consciousness. A representational theory of the phenomenal mind*. London: The MIT Press.

Velleman, J. D. (2000). *The possibility of practical reason*. Oxford: Oxford University Press.

---

[1] Ver por ejemplo: Feigl (1967), Rorty (1979).

[2] En palabras de Tye (1995), esa “cualidad experiencial distintiva” que no puede faltar en ningún estado consciente (p. 3). Para Block (1995), “la consciencia-F [la consciencia fenoménica] es experiencia. Las propiedades conscientes-F son experienciales. Los estados conscientes-F son experienciales, esto es, un estado es consciente-F si tiene propiedades experienciales” (p. 230).

[3] Hay también quienes niegan que haya un concepto de mente de sentido común, no teórico (véase, por ejemplo, Rorty 1979). No puedo discutir aquí este asunto, sin embargo.

[4] Para caracterizaciones de la inteligencia en esta línea, véase James 1950; MacIntyrey Nowell-Smith, 1960; Ryle, 1963.

[5] Para una noción de intencionalidad en esta línea, que no excluye situaciones como la mencionada, cf. Tye, 1995.

CAPÍTULO II

La conciencia en la discusión actual: cuestiones filosóficas

Marcelo Díaz

Introducción

Una manera de plantear la diferencia entre el análisis conceptual (filosófico) y la investigación empírica científica es preguntándonos sobre lo que tendríamos que hacer para responder a una pregunta filosófica, a una pregunta científica y a otra de sentido común (Díaz, 2012). Examine las siguientes preguntas:

1. ¿Qué hora es?
2. ¿Qué función cumplen los péptidos en el cerebro?
3. ¿Cuál es la distribución geográfica de las etnias en los Estados Unidos?
4. ¿Cuál es la raíz cúbica de 729?
5. ¿Qué es la verdad?
6. Juan está tratándose con un neurólogo. Juan suscribe la tesis de la identidad entre la mente y el cerebro. De su médico ha recibido la noticia que le han localizado un tumor cerebral y que es urgente extirparlo. ¿Creerá Juan que la operación le va a amputar parte de su mente?
7. ¿Qué es la filosofía?
8. ¿Existe vida en el planeta más alejado de nuestra galaxia?
9. ¿Es la indeterminación un rasgo esencial del mundo subatómico?
10. ¿Qué es la conciencia?

Ahora bien, hay preguntas que se responden recurriendo a la experiencia cotidiana que incluye la participación de órganos sensoriales. Este parece ser el caso de la pregunta 1; para responder otras, las preguntas 2 y 3, se necesita información obtenida en el marco de teorías que guían la experiencia del investigador del cerebro o del etnógrafo, respectivamente. La pregunta 4, en cambio, se contesta realizando cálculos o, en términos generales, siguiendo algún método deductivo, ateniéndose a reglas formales. La pregunta 5, por su parte, nos deja perplejos, sentimos que no hay nada que podamos indicar o mostrar, se trata de una pregunta filosófica; es decir, no es una pregunta acerca de los hechos o estados de cosas en el mundo, sino acerca de los conceptos en términos de los cuales comprendemos el mundo y comunicamos tal comprensión. En otras palabras, son preguntas acerca de nuestros marcos conceptuales. Estos conceptos tienen la peculiaridad de ser muy generales; piénsese por ejemplo en el concepto de ser, de tiempo, de verdad, de materia o de espíritu. Podríamos decir que estos conceptos son como lentes a través de las cuales "observamos" el mundo.

Si bien las preguntas filosóficas o conceptuales adoptan habitualmente la forma típicamente socrática "¿qué es X?" (donde X es la verdad, la materia, el tiempo, el espíritu, etc.), a veces adoptan otras modalidades que hacen que fácilmente puedan confundirse con preguntas científicas o cotidianas. Algo de esto es lo que encontramos en las preguntas 6 y 9. Para responder la pregunta 8 cabe precisar qué significa que hoy estemos en condiciones de especificar si hay o no vida en algún lugar por muy remoto que sea, y distinguir entre una confirmación actual y una en principio. Esta pregunta es polémica porque podría ser una genuina pregunta filosófica o conceptual, pero también podría darse buenas razones para considerarla científica o incluso de ciencia ficción. La pregunta 10 plantea la cuestión de la naturaleza de la conciencia, qué lugar o puesto tiene en el cosmos. Esta pregunta nos desconcierta debido a que entendemos por "conciencia" o "estar (o ser) consciente" una diversidad de situaciones de ser; esto es, propiedades, complejos de propiedades, etc.

A modo de introducción, pensemos primero en otro concepto con el cual está emparentado, nos referimos al concepto de "espíritu" (Campbell, 1987) o, si somos generosos, al concepto de "mente". Si nos preguntamos a propósito de este concepto (el de espíritu o el de mente) qué condiciones

deberían darse para que algo sea candidato a objeto espiritual o mental, una de las respuestas sería que debe ser algo que posea conciencia o la propiedad de ser o estar consciente, tal como nos lo enseñó Descartes. La otra opción que caracteriza a un objeto espiritual o mental según una venerable tradición filosófica y psicológica es la intencionalidad, propiedad que, según Brentano, radica en que los estados mentales están dirigidos a algo (un deseo, por ejemplo, es un deseo de algo: el deseo de tomar un helado; un recuerdo es el recuerdo de algo: el recuerdo de lo que comí en la última cena de Navidad, etc.). En este sentido, podríamos decir que ambos conceptos, los de "conciencia" e "intencionalidad", están conectados con el concepto de "espíritu" o de "mente", por más que a este último se le omita en las discusiones filosóficas actuales, tal como suele hacerse con el concepto de "materia", concepto que asociamos de inmediato cuando nos preguntamos acerca de la filosofía que subyace en la mayoría de los enfoques científicos de la conciencia.

Es innegable que tenemos buenas razones para creer que la conciencia, en caso de poder incluirse en una concepción científica del mundo, debe ser algo material, porque la ciencia empírica (natural o social) tiene que ver con objetos o sistemas materiales. Pero de inmediato surgen nuevas interrogantes: ¿qué es la materia?, ¿puede reducirse la mente y con ello la conciencia a algo material como el cerebro y sus propiedades bioneurales? Cuando hablamos de la base material de la conciencia, ¿puede entenderse por ello ya no un cerebro vivo sino un artefacto tecnológico como el computador? Veamos a continuación lo que podemos extraer de un análisis conceptual (usos lingüísticos) de la conciencia.

Análisis conceptual de la conciencia

En lo que concierne al concepto de "conciencia" no es ocioso preguntarse sobre los usos que se le suele dar. "Conciencia" o "ser (o estar) consciente" se puede emplear en diversos sentidos. Veamos algunos de estos usos tanto cotidianos como técnicos (Díaz, 2013):

(a) Cuando decimos de un animal o de una persona que está consciente porque se encuentra en estado de vigilia, o porque es receptivo a las señales de su entorno (su opuesto sería estar dormido, anestesiado, en coma o

inconsciente).

(b) Cuando lo usamos para indicar que alguien ha alcanzado la etapa de su madurez en que es capaz de poner atención a las cualidades de su carácter o de su intelecto. En este caso su opuesto se conecta con el uso freudiano de “inconsciente”, es decir, aquel estado en que la persona no es capaz de reconocer sus falencias, errores, fobias, celos o impulsos eróticos porque sus apreciaciones revelan determinados prejuicios o porque es inmadura.

(c) Cuando se quiere decir que una persona está prestando atención a algo, por ejemplo, a la mosca que revoletea sobre su cabeza mientras lee este texto.

(d) Cuando se usa como sinónimo de “descubrir” o “darse cuenta de”. Por ejemplo, el profesor se dio cuenta (tomó conciencia) de que hay menos alumnos de lo habitual en la sala de clases.

(e) Cuando “conciencia” alude a la introspección o reflexión, esto es, cuando deliberadamente inspeccionamos nuestros pensamientos o, ya en el contexto de la psicología, de la lingüística o de la inteligencia artificial, cuando tenemos pensamientos de segundo orden acerca del hecho de estar en un determinado estado mental (metacognición).

(f) Cuando “conciencia” quiere decir “autoconciencia” o identidad personal, es decir, la posesión de un concepto de la propia unidad en el tiempo de nuestros diversos estados mentales.

(g) Cuando significa conciencia de acceso –tal como lo propone Ned Block y como recoge un *Diccionario de ciencias cognitivas*–, cuando un estado es consciente “si, en virtud del hecho que alguien se halle en ese estado, una representación de su contenido se hace inmediatamente disponible como premisa para el razonamiento, y puede intervenir en el control racional de la acción y de la palabra” (2003, p. 92).

(h) Cuando “ser (o estar) consciente” quiere decir “intencionalidad”, es decir, que nuestros estados mentales están dirigido a algo. En este caso estamos entendiendo la conciencia como una relación: la conciencia es siempre conciencia de algo (Brentano)[1].

(i) Cuando “conciencia” quiere decir la manera como se nos presentan subjetivamente nuestras sensaciones (y quizá la totalidad de nuestros estados mentales). Esta noción es heredera de otras nociones y/o doctrinas como los *sense-data* de los filósofos empiristas y/o fenomenalistas, los

sensibles propios de Aristóteles, etc.

El último caso es el más problemático y es al que aluden los filósofos cuando hablan de *qualia* (plural) o *quale* (singular). Se trata de la conciencia fenoménica, el cómo es tener una determinada experiencia, la propiedad cualitativa y subjetiva de la experiencia consciente. Pero el problema de la conciencia y sus *qualia* es del tipo que se resiste a ser abordado directamente; parece que debemos dar un rodeo. Lo primero es despejar la muy constante confusión entre la conciencia y sus *qualia*, por un lado, y las llamadas cualidades secundarias por el otro. La doctrina de que hay cualidades secundarias diferenciadas de las cualidades primarias se debe, entre otros, a John Locke y Robert Boyle (Aristóteles se había adelantado al diferenciar entre sensibles comunes y sensibles propios). Las cualidades primarias forman parte del mundo físico en el sentido de que son las cualidades mensurables de las que se ocupa la ciencia (peso, longitud, masa, figura o forma). No son propiedades existentes por referencia a los seres humanos. Las secundarias, en cambio, aluden a los poderes, capacidades o disposiciones de los objetos materiales, de producir en los sujetos perceptores, gracias a sus cualidades primarias, determinadas experiencias sensoriales, tales como el olor, el sabor, el color, el sonido y el calor o el frío. Tienen una fuerte carga subjetiva o antropológica, pero su punto de arranque está en el mundo físico; no son propiedades de la experiencia de los objetos, sino propiedades de los objetos de experiencia.

Los *qualia*, por el contrario, serían exclusivamente rasgos de nuestra experiencia: la rojez del rojo, lo doloroso del dolor, etc. Los *qualia* serían propiedades o rasgos asociados a la experiencia consciente; es decir, no hay *qualia* inconscientes. Podrá concederse la posibilidad de cualidades secundarias que percibamos sin ser consciente de ellas, pero la experiencia del color que tengo cuando efectivamente me apercibo de ello, sería la experiencia de un *quale*.

Pero la conciencia y sus *qualia* han presentado enormes desafíos a los científicos, tanto en la investigación empírica como en la modelización de la misma. Varias obras científicas de las últimas décadas han generado una sensación optimista en cuanto a los logros obtenidos, especialmente en la explicación neurofisiológica de la conciencia visual, pero al mismo tiempo se ha ido aceptando que hay dificultades enormes y quizá insalvables en abordar y esclarecer la conciencia fenoménica, esto es, las cualidades

subjetivas de la experiencia consciente. Todo ello ha configurado lo que los neurocientíficos llaman “el problema duro” (Aboitiz, 2001).

Por otro lado, algunos filósofos han pensado que el carácter consciente de un estado mental se explica en términos de una particular relación de tipo cognitivo que se establece entre ese estado y otro del mismo sistema, en virtud de lo cual el primero adquiere la propiedad de ser consciente (Stigol, 1998). Si dicha relación no existe, entonces el estado mental no será consciente. Tal relación es concebida como percepción interna (Armstrong) o como pensamiento de segundo orden (Rosenthal). A estas teorías se las puede llamar “teorías del monitoreo interno” porque tienen en común:

(a) La propuesta de una definición relacional de la conciencia que opera bajo la suposición que nuestra mente es capaz de reflexionar sobre sí misma, que es capaz de desdoblarse en el sentido de que un estado de ella –estado de segundo orden– puede dirigirse intencionalmente a otro estado de ella misma, uno de primer orden.

(b) La tesis de que ambos estados tienen una existencia distinta e independiente, de modo que es posible que uno de ellos ocurra sin que ocurra el otro. Solo contingentemente uno acompaña al otro, de modo tal que la conciencia resulta ser un rasgo extrínseco que de modo accidental poseen algunos estados mentales.

Entre las objeciones generales que es posible hacer a estas teorías se señala que, más que explicar la conciencia misma –la conciencia entendida por la propiedad de la subjetividad–, aparentemente se ha ofrecido una explicación del autoconocimiento o de la introspección.

Es en este punto que quisiera introducir la figura de Ludwig Wittgenstein, dejando de lado la contribución a la discusión de famosos pensadores como John Searle, Daniel Dennett, David Chalmers, Roger Penrose y tantos otros. Es un lugar común citar las *Philosophical Investigations* del pensador austriaco cuando se quiere saber acerca de cómo él enfrentó desde el análisis del lenguaje los temas que hoy rotulamos como de filosofía de la mente. Sin embargo en las últimas décadas hemos descubierto una verdadera cantera filosófica en sus obras que no fueron libros propiamente tales, más bien fueron notas dictadas a sus colegas o discípulos, pensamientos sueltos, apuntes de clases de sus estudiantes. Usaré como guía justamente los análisis que llevó a cabo Wittgenstein frente a sus alumnos de Cambridge en la segunda parte del año académico

1935-1936 (*Philosophical Ocassions 1912-1951*, 1993; versión en español titulada como *Ocasiones Filosóficas 1912-1951*, 1997). En ocasiones nos apoyaremos en sus *Philosophical Investigations.* Nuestra tesis es que tales análisis son perfectamente contemporáneos con momentos estelares de las actuales ciencias cognitivas y también, por supuesto, contemporáneos de las discusiones que son propias de la filosofía de la mente.

Wittgenstein, experimentos mentales y datos de los sentidos

Como vimos anteriormente, un tema muy debatido en las ciencias cognitivas y en la filosofía de la mente es el de la conciencia y sus *qualia*, es decir la conciencia en su aspecto estrictamente subjetivo. Uno de los pensadores que contribuyó decisivamente a familiarizarnos con este tema fue John Locke a través de un caso de inversión cromática entre dos personas (inversión cromática interpersonal). Se trata del caso en que la idea de azul fuese diferente entre dos personas. Intentando explicar esa situación escribe en “El ensayo sobre el entendimiento humano”:

> Tampoco podría imputarse a nuestras ideas simples, si por la diferente estructura de nuestros órganos estuviese ordenado que un mismo objeto produjera, al mismo tiempo, diferentes ideas en las mentes de diversos hombres. Por ejemplo, que la idea que produjera una violeta en la mente de un hombre por conducto de su vista fuese la misma idea producida en la mente de otro hombre por una caléndula y viceversa (Locke, 1956, p. 375).

La hipótesis del espectro invertido de Locke chocaba con el ideario programático del Círculo de Viena al resultar inaplicable el principio de verificabilidad empírica, la exigencia de poder verificar, al menos en principio y de manera intersubjetiva, si una persona sufre o no la inversión cromática. Como sabemos en su primera versión el Círculo de Viena suscribía un fenomenalismo que impedía acceder directamente a los datos sensoriales de otra persona, de manera que el problema se rotuló como un típico caso de pseudoproblema[2].

Wittgenstein por su parte, sin embargo, sigue a Locke con un ejemplo de inversión interpersonal al inicio de su clase del 26 de febrero de 1936 en los términos siguientes:

> Quiero hablar sobre la idea de privacidad de los datos de los sentidos. La idea surge del hecho de

> que otro hombre y yo podríamos mirar a algo y yo podría decir “él lo ve de un color distinto del que yo lo veo”. Podría hacer esto en el caso de que uno de nosotros llevara puestas unas gafas coloreadas. O supóngase que le enseño a una persona los nombres de los colores bajo una luz determinada. Después él es capaz de traerme los objetos rojos que hay en la habitación cuando se los pido. Supóngase que entonces le llevo a la luz del día, y que llama a los pétalos de rosa verdes y a la hierba roja, y así sucesivamente. Si lo llevamos después otra vez a la luz artificial y se comporta como antes, entonces podríamos decir: “Ve rojo lo que nosotros vemos verde”. Aquí, los criterios para que yo diga: “Ve verde lo que yo veo rojo” son muy complicados, aunque el significado es muy simple” (Wittgenstein, 1997, p. 281).

Este es un ejemplo de inversión de colores interpersonal, algo más complejo que el ejemplo de Locke. La noción de idea en Locke se estaba empleando, en este contexto, en su sentido de imagen mental subjetiva. Aún no se realizaba la proeza epistemológica de reemplazar la estrategia de las ideas tan explotada por él por una estrategia semántica consistente en discutir acerca de las palabras que recogen nuestra experiencia. Con Wittgenstein se introduce la noción de criterio, problemática como el mismo reconoce porque la cuestión es: ¿con qué criterio digo de otra persona que él ve algo de un color distinto al que yo veo? La cuestión que está planteando es difícil y clásica, es la cuestión de la privacidad de los datos sensoriales, en particular los colores. Dicho de una manera breve, el asunto que se tiene entre manos Wittgenstein es cómo mediante un recurso no privado puedo hacer un seguimiento de algo que es, al parecer, intrínsecamente privado.

¿Qué sucedería, plantea Wittgenstein, si un ciego que recupera la visión dijera: “Antes, cuando alguien me golpeaba, veía lo que ahora denomino luces” (1997, p. 282). Si admitiéramos la posibilidad de que realmente siempre pudo ver pero que su comportamiento era diferente, porque por ejemplo a diferencia nuestra chocaba con los objetos, no podríamos decir, en sentido estricto, que “sabíamos” que él siempre vio lo que nosotros vimos; en realidad deberíamos ser cautos y decir que más bien solo sabemos cómo se comporta, pero que no tenemos idea de si él veía o no veía. Es decir hay casos en lo que no sabemos algo pero sí podemos conjeturar que si esa persona se comporta como todos nosotros, se sigue que esa persona probablemente ve. Hay aquí algo interesante: Wittgenstein está apuntando a que nuestro acceso a la experiencia visual de otra persona

es más bien un acceso indirecto: partimos con la observación de su comportamiento y luego conjeturamos diciendo “se comporta como yo, luego probablemente ve” (Wittgenstein, 1997, p. 282). En estas líneas –o mejor, palabras emitidas por Wittgenstein en su clase– él tácitamente está admitiendo que adjudicar una determinada experiencia a otra persona tiene ciertas peculiaridades: tenemos que usar como premisa (esté o no oculta) la afirmación de que uno tiene un acceso directo a lo que le está sucediendo, en este caso, a la propia experiencia de ver y que, en cierto sentido, uno sabe que está viendo; en cambio en el caso de la experiencia visual de otro solamente podemos conjeturarlo. Uno puede decir “probablemente Watson ve” y esto es una conjetura, pero para que sea efectivamente una conjetura hay que tener presente el uso que se hace de la expresión porque podría darse la situación en que uno diga “probablemente Watson ve” y lo que uno esté haciendo es una afirmación retórica sobre una cuestión de hecho. Hay que fijarse, nos advierte Wittgenstein, en que para distinguir entre una conjetura propiamente tal y una afirmación retórica, debemos considerar también el tono de voz con que se dice la oración incluyendo las de las palabras “quizás” o “probablemente”.

Wittgenstein parece descartar una salida radicalmente escéptica del tipo “nunca puedo saber si él ve o no”. Pero se allana a considerar un caso puntual de escepticismo, cuando la duda es acerca de un *quale* determinado, un color determinado, el rojo, esto es que “nunca sabemos si ve rojo lo que nosotros vemos rojo”. Wittgenstein hace la diferencia fina entre este caso que alude a una cuestión epistémica y el caso diferente en que uno podría decir “por supuesto, sabemos si entiende por la palabra ‘color’ lo mismo que nosotros” (Wittgenstein, 1997, p. 282). Se ha deslizado casi imperceptiblemente, de lo epistémico a lo lingüístico, a las palabras. La pregunta obvia es qué gana con tal cambio. La respuesta creo que no puede ser otra que la siguiente: se gana en manejo del asunto, no es lo mismo discutir acerca de un *quale* determinado con toda la carga subjetiva que conlleva y discutir acerca de una palabra, esto es de una entidad pública por más que lo que ella mienta sea algo privado.

El análisis de Wittgenstein se empieza complejizar cuando se refiere a las impresiones que uno puede tener. ¿Qué quiere decir que uno tiene una impresión?, ¿puedo acaso ponerle un nombre? Veamos lo que les dijo a sus alumnos: “Voy a dar un nombre a esta impresión, por ejemplo, el nombre

‘alfa’, y entonces sé lo que significa el nombre ‘alfa’” (Wittgenstein, 1997, p. 283). Aquí Wittgenstein está introduciendo un tema elaborado cuidadosamente en sus *Philosophical Investigations,* la cuestión del nombrar como función del lenguaje. Al respecto en su clase del 26 de febrero de 1936 dijo lo siguiente:

> Supóngase que se le da un nombre a una persona. ¿Existe una diferencia entre poner una raya azul en su frente y poner una raya que leo como “Watson”? ¿Es poner una raya simplemente dar un nombre? Se podría responder que sí o que no, dependiendo de lo que se quiera decir. Una palabra que nosotros conectamos con una cosa puede tener todo tipo de funciones diferentes. El nombre de un diamante, Koh-i-noor, no tiene la misma función que el nombre de una persona. Puedo *llamar* a una persona por su nombre, pero no puedo hacer eso con el nombre de un diamante. Para dar nombre a una persona he de dar un rasguño *conectado* con las convenciones para su uso (Wittgenstein, 1997, p. 283).

Resulta inevitable reconocer en estas líneas toda su diatriba con la tradición que representa San Agustín en sus *Confesiones* y que incluye por extensión, como sabemos, la crítica a pensadores contemporáneos de Wittgenstein y también su autocrítica. Recordemos un pasaje muy citado de las *Investigaciones filosóficas*:

> Para una gran clase de casos de utilización de la palabra “significado” –aunque no para todos los casos de su utilización– puede explicarse esta palabra así: el significado de una palabra es su uso en el lenguaje. Y el significado de un nombre se explica a veces señalando a su portador (1988, p. 61).

De acuerdo con el texto (parágrafo 43) sería descaminado preguntar por el significado de las palabras, lo correcto sería inquirir por sus usos. Pero también y en perfecta coherencia con el texto hay un caso excepcional: los nombres. ¿Qué quiere decir esto? Veamos más en detalle este punto.

Hay casos en los que el significado de las palabras se explica señalando ostensivamente el objeto portador del nombre. Esta explicación extendida a todo el funcionamiento del lenguaje tiene, como acertadamente reconoce Wittgenstein, una venerable tradición. El ya mencionado san Agustín de Hipona, pero también el neopositivista Rudolf Carnap, su maestro Bertrand Russell y él mismo en la etapa inicial de su filosofía, la del *Tractatus Logico-Philosophicus*, creyeron que la función esencial del lenguaje consiste en nombrar y representar el mundo: “Nombrar algo es similar a

fijar un rótulo en una cosa" (Wittgenstein, 1988, p. 29). La semejanza con la cita de sus lecciones en Cambridge es evidente.

Pero ¿qué ocurre cuando el nombrar se da en el contexto de una impresión tal como lo planteó a sus alumnos? La pregunta no es retórica porque plantea la cuestión de cómo hay que concebir el lenguaje de nuestras experiencias, de nuestras sensaciones. Es decir, se está inquiriendo sobre lo que ocurre cuando bautizamos una impresión. Veamos una vez más una cita de *Ocasiones filosóficas 1912-1951* sobre el trasfondo de sus *Investigaciones filosóficas*:

> Ahora bien, si llamo "alfa" a una *impresión* determinada, parece como si hubiera realizado un acto de bautismo. Pero lo que pudiera haber pasado aquí al darlo no importa mucho. Por supuesto, tengo un estado mental determinado. Miro fijamente a la impresión. Pero parece que simplemente he mirado fijamente y he pronunciado la palabra "alfa". No sabemos en absoluto qué hacer con la palabra, y en consecuencia no puede decirse en absoluto que tenga un significado (Wittgenstein, 1997, p. 283).

Aquí Wittgenstein está ventilando la cuestión de lo que implica suponer que al dar un nombre a una impresión –o a una sensación determinada– uno le está dando significado al signo (nombre) mediante una especie de definición ostensiva. Tal como lo plantea en el parágrafo 258 del diarista de las *Philosophical Investigations,* tal definición ostensiva podría ser como un acto en que uno señala internamente la sensación concentrándose. Pero todo esto no es más que una ceremonia, lo extraño es que se espera que mediante una definición uno pudiera establecer el significado del signo. Pero qué quiere decir en este caso que uno se imprime ostensivamente la conexión del signo con la sensación:

> "Me la imprimo", no obstante, solo puede querer decir: este proceso hace que yo me acuerde en el futuro de la conexión correcta. Pero en nuestro caso yo no tengo criterio alguno de corrección. Se querría decir aquí: es correcto lo que en cualquier caso me parezca correcto. Y eso solo quiere decir que aquí no puede hablarse de "correcto" (Wittgenstein, 1988, p. 227).

En su clase del 3 de marzo de 1936 vuelve sobre el asunto de la privacidad de los datos sensoriales, pero introduce una nueva variable que enriquece aún más su análisis: el imaginar que otra persona ve un color determinado, que ve verde lo que yo veo rojo. ¿Qué significa imaginar en este caso? Una respuesta sería que la otra persona me pinte con los ojos

abiertos y con un objeto rojo delante de mí. Sin duda que tal situación puede darse, sin embargo Wittgenstein advierte lo siguiente: “Imaginar que yo miro a un objeto rojo es algo fácil. Imaginar que yo veo rojo es difícil” (Wittgenstein, 1997, p. 286). Nuevamente nos advierte que hay algo, un residuo claramente subjetivo que no se cubre cuando se pinta a una persona con los ojos abiertos y con objetos delante de él. Como veremos más adelante es en este tipo de casos analizados por Wittgenstein en los que la tradición cartesiana se hace fuertemente presente como un gran desafío a superar. Dejemos momentáneamente nuestro comentario de Wittgenstein y agreguemos un caso de nuestra propia factura.

Inversión cromática interpersonal, el imaginar y la memoria

Considérese la siguiente situación: “Luis es un reportero de guerra que ha tenido un percance, ha recibido una bala de bajo calibre que entró y salió de su cabeza. Estuvo hospitalizado y le dieron de alta médica. Ha regresado a su país y va de compras con su novia a una feria de abastos. El día se anuncia esplendoroso y él mirando al cielo le dice a su pareja: ‘¡Vamos a tener un día completamente despejado, mira el cielo, está de un rojo intenso!’. Más tarde al ver unos tomates maduros como a él le gustan, suelta la siguiente oración: ‘¡Esos que están ahí, los tomates de color azul, esos quiero comprar!’. Su novia está sorprendida y pensando que Luis deberá ir pronto a un neurólogo o a un psiquiatra, le pide que no hable de esa manera extravagante y que diga que el cielo despejado es de color azul y los tomates maduros de color rojo. Luis asiente y comienza a hablar del color del cielo despejado y de los tomates maduros en los mismos términos que los demás y que él mismo antes que la bala haya cambiado dramáticamente su percepción de los colores”.

Fíjense que si bien aparentemente Luis está finalmente usando los términos “azul” y “rojo” con el mismo carácter representacional o intencional que las demás personas, hay algo que en Luis ha quedado completamente alterado por más que él hable del cielo despejado azul y los tomates maduros rojos. Su vivencia, el carácter estrictamente subjetivo de su experiencia cromática se ha invertido. Este experimento mental es una descripción a lo que nos tienen acostumbrados los filósofos que desean

convencernos que hay tal cosa como los *qualia* (o *quale),* es decir la experiencia subjetiva de ver algo con un color determinado y que no puede confundirse con el contenido intencional o representacional de nuestros estados mentales.

Pero con independencia de si se trata de un caso de inversión de colores interpersonal o intrapersonal, hay algo que está en la base de estas discusiones: que se está dando por sentado que hay efectivamente un problema consistente en justificar nuestra creencia en que los demás son personas con deseos, pensamientos, recuerdos, etc. y no autómatas. ¿Cómo es que se instaló la necesidad de justificar lo que en la vida cotidiana nos parece completamente aproblemático, transparente? Esto nos lleva a Descartes, asunto que por falta de espacio quisiera resumir en un par de líneas: Descartes se vio en la necesidad de justificar racionalmente su creencia en que no está rodeado de autómatas sino que por personas debido a la propia lógica interna de su dualismo psico-físico. Este dualismo de substancias desembocó en una distinción tajante en el plano epistémico: habría una diferencia epistemológica fundamental entre acceder a las operaciones de nuestra propia mente y acceder a las operaciones mentales de otros.

Ahora se traerá a colación una observación del filósofo Alfred Ayer y se verá cómo Wittgenstein aborda el problema mencionado por el primero. Dice Ayer: “Hay un sentido correcto y familiar en el que se puede decir que dos personas diferentes perciben el mismo objeto, oyen el mismo sonido, sienten la misma sensación, y de esto se sigue que tienen las mismas experiencias” (1979, p. 179).

Puede agregarse a lo anterior que aunque esas dos personas perciban los mismos objetos o escuchen los mismos sonidos, no sienten los mismos datos sensoriales porque estos, los datos sensoriales, son privados por definición. Como lo dijese el mismo Ayer:

> [E]stán caracterizados de tal manera que la proposición de que una persona tiene los datos sensoriales de otra no describe ninguna situación posible. De manera similar, si una persona tiene miedo y otra comparte su sentimiento, se dirá también que hay dos sentimientos de miedo y no uno solo. Decir que una sensación es compartida equivale a decir que dos sensaciones son cualitativamente similares y que tienen el mismo objeto ostensible: no es decir que sean numéricamente idénticas. Pero, ¿cómo están diferenciadas? No se puede suponer que se pueden

> contar las sensaciones de las personas como se pueden contar las cosas que puedan llevar en sus bolsillos. La respuesta es que tenemos que decir que hay dos sensaciones y no una sola, justamente porque hay dos personas. Se establece un acuerdo de que toda sensación que uno tiene es una experiencia que es privativa de uno mismo. Y por tanto se convierte en una verdad necesaria que una persona no puede tener, y por tanto no puede conocer en sentido estricto, las experiencias de otros (Ayer, 1979, p. 179).

Los dos pasajes del ensayo de Ayer recién citados tiene el mérito de ponernos de frente a la cuestión que Wittgenstein abordó una y otra vez en sus escritos y en sus clases: la cuestión de la privacidad de nuestra experiencia y del lenguaje que da cuenta de tal privacidad. Este es un problema que nos remite a la tradición cartesiana que tanto enfatiza el carácter no público sino privado, interno y subjetivo de nuestra experiencia, al punto que se transforma en un problema –tal como ya se dijo anteriormente– el tener que justificar la creencia de que hay cuerpos fuera del mío que expresan los pensamientos de mentes ajenas a las cuales no puedo acceder directamente. Wittgenstein en su clase del 3 de marzo de 1936 comunicó a sus discípulos lo siguiente:

> No hay ninguna objeción a la adopción de un modo de expresión en el que decimos: “Miró al cielo, tuvo un dato sensorial, y dijo ‘El cielo está azul, vámonos de paseo’. O también: “Tuvo un dato sensorial, dijo ‘Vete al infierno’”, etc. ¿Pero cómo se usa en este caso “tuvo un dato sensorial”? Podemos imaginar lo que significa decir “tuvo un dato sensorial”, pero en muchos casos no tiene ningún uso (Wittgenstein, 1997, p. 287).

Que aquí se está planteando la cuestión fundamental de la privacidad de nuestra experiencia y del lenguaje que da cuenta de ello es algo que queda patente cuando nos dice a continuación:

> “Sé cómo es que yo imagine rojo”.
>
> Bien, ¿cómo?
>
> “No puedo explicárselo a otra persona, pero puedo explicármelo a mí mismo. Lo hago imaginando rojo” (Wittgenstein, 1997, p. 287).

Nuestro comentario es el siguiente: cabe preguntarse si esto es o no una explicación porque cabe preguntar por el criterio que se está empleando para decir que esto es rojo ahora. Una posible respuesta, no exenta de dificultades, sería recurrir a la memoria como el criterio que me permite

decir que ahora recuerdo haber dado ese nombre a este color con anterioridad. Es claro que la memoria suele fallar pero aquí el error de la memoria parece estar descartado *a fortiori* pues estamos hablando de un lenguaje privado, de un lenguaje en el que los nombres de colores estarían directamente vinculados o conectados ostensivamente con la experiencia del color (como en esa especie de ceremonia típicamente agustiniana). Es por esto que Wittgenstein crípticamente señala que "no hay posibilidad de que la memoria me haga una jugarreta, porque (en este caso) no puede haber un criterio para que me haga una jugarreta" (Wittgenstein, 1997, p. 287).

¿Cuándo entonces la memoria me puede hacer una jugarreta? La respuesta es que eso ocurre cuando se trata de, por ejemplo, recordar el título de un libro, un número telefónico o el nombre de una canción. En cambio, en el caso en que se esgrime la memoria como el criterio que nos permite respondernos a nosotros mismos si estamos asignando o no correctamente el término "rojo" a lo que estamos ahora experimentando, nos obliga a no salirnos del ámbito estrictamente privado. Recordemos que interpretando el famoso parágrafo 258 del diarista en sus *Philosophical Investigations*[3]: si el signo "S designa o nombra una sensación recurrente, se sigue que el criterio o regla que indica cuando debe usarse la oración "esta es la sensación S" o cuando la oración "esta no es la sensación S" es algo privado, *viz*., la memoria. Pero ¿es la memoria totalmente confiable?

En la vida cotidiana nos suelen asaltar dudas acerca de nuestros recuerdos (lo que Wittgenstein llama "jugarretas que nos hace la memoria"). Por ejemplo, nos preguntamos: ¿Cuál es el título del libro cuya reseña leímos en un periódico hace unas semanas atrás?, ¿cómo se llama la última canción que compuso Violeta Parra?, ¿cuál es el número telefónico de nuestro mejor amigo de infancia?, etc. En todos estos casos nuestras dudas pueden disiparse mediante información objetiva que nos permita determinar si nuestra memoria falla o acierta. Encontrar el periódico, consultar una biografía de la autora de *Gracias a la vida*, consultar una vieja agenda, etc., serían los criterios que necesitaríamos invocar. Pero cuando se trata de recordar una sensación estrictamente privada no tenemos nada objetivo que nos permita precisar si la memoria falla o acierta. Si un individuo dijera que le parece o que le da la impresión o, para mencionar una expresión popular, que le tinca que la sensación que está

experimentando es la sensación S, no habrá manera de decidir con precisión si es o no es efectivamente así.

Acá, el impedimento no es empírico: si lo fuera podríamos, en principio, solucionar la dificultad, como ocurre cuando deseamos recordar el nombre de un libro, de una canción o de un número telefónico. En el caso de un suceso entera o esencialmente privado, como se supone que lo es una sensación, no podemos apelar a ningún criterio objetivo para establecer comparaciones entre las sensaciones pasadas y la actual. Estamos forzados por la manera de concebir la función del signo "S" a no salirnos del ámbito privado. Estamos sumidos en un radical egocentrismo. En cambio, cuando seguimos una regla nos involucramos en un conjunto de actividades porque tales reglas y acciones, junto con las palabras, forman el juego lingüístico. Es un contrasentido, por lo tanto, suponer que el signo "S" pueda estar regulado en su uso por una regla privada: dicha privacidad sería un rasgo incompatible con la noción de regla. Insistir en la suposición de que hay reglas privadas sería, para usar un ejemplo del mismo Wittgenstein, como el caso de un traductor que trabaja con un diccionario imaginario. Si nos preguntáramos ¿ha traducido correctamente o incorrectamente?, la respuesta estaría bloqueada, porque para obtener la respuesta tendríamos que tener acceso a ese diccionario imaginario, el cual *ex hypothesi* nos está vedado.

Conclusiones

Finalmente quiero destacar que en la postura de Wittgenstein, además de encontrarse una crítica a los enfoques fenomenalistas bajo la forma de un análisis del lenguaje de los datos sensoriales, hay una crítica de mayor alcance. Esta es la crítica a la tradición cartesiana. Es el cuestionamiento que él hace a la tendencia a ver las palabras acerca de nuestra propia experiencia como si adquiriesen su significado designando ostensivamente desde la experiencia misma, el proceso que ocurre en un ámbito que podemos llamar "el teatro cartesiano" (Daniel Dennett), "el fantasma en la maquinaria corporal" (Gilbert Ryle), o simplemente el ámbito de nuestro mundo interno. De estos hábitos lingüísticos se han seguido una serie de problemas en apariencia irresolubles y tachados como problemas

tradicionales de la filosofía, entonces es útil seguir a Wittgenstein en aquello de que la filosofía es “la lucha contra la fascinación que ejercen sobre nosotros las formas de expresión” (*Los cuadernos azul y marrón*, p. 56).

Agradecimientos

Este trabajo fue escrito en el contexto del proyecto Dicyt (Usach) Código 031853DS (2018-2019), Universidad de Santiago de Chile.

Referencias

Aboitiz, F. (2001). “Sincronía, conciencia y el problema duro de la neurociencia”. *Revista Chilena de Neuro-Psiquiatría*, *39* (40), 281-285.

Ayer, A. (1979). El conocimiento de otras mentes. *Ensayos filosóficos.* Barcelona: Ariel.

Campbell, K. (1987). *Cuerpo y Mente*. México: Unam.

Díaz, M. (2010). Sobre el lenguaje y las sensaciones en la filosofía de Wittgenstein. En S. Rivera y A. Tomasini (Eds.), *Wittgenstein en Español II* (pp. 99-121). Buenos Aires, Argentina: Ediciones de la Universidad Nacional de Lanús.

——(2012). La conciencia desde un punto de vista filosófico. Cuestiones conceptuales. *Revista Chilena de Neuropsicología*. 7(1), 7-11.

Díaz, M. (2013). El materialismo en la filosofía de la mente y en las ciencias cognitivas. *Scripta Philosophiae Naturalis*, 4 (Juillet-Décembre), 43–57.

Houdé, O. *et al.,* (2003). *Diccionario de ciencias cognitivas.* Buenos Aires: Amorrortu Editores.

Locke, J. *(1956). Ensayo sobre el entendimiento humano*. México: Fondo de Cultura Económica.

Stigol, N. (1998). “Conciencia fenoménica y acceso cognitivo”. *Análisis Filosófico*, vol. XVIII (2), 131-141.

Wittgenstein, L. (1988). *Investigaciones filosóficas*. Barcelona: Editorial Crítica.

—— (1989). *Los cuadernos azul y marrón.* Madrid: Tecnos.

—— (1997). *Ocasiones filosóficas 1912-1951*. Madrid: Cátedra.

---

[1] En la actualidad en filosofía de la mente se distingue claramente entre conciencia e intencionalidad

en los términos respectivos del aspecto cualitativo o subjetivo de los estados mentales, por un lado, y su aspecto relacional o intencional por el otro.

[2] Para más detalles sobre el origen del experimento pensado de la inversión intrapersonal véase B. Gert (1965) Imagination and verifiability", *Philosophical Studies*16, 44-47; H. Putnam (1965) Brains and behavior" en R.J. Butler (comp.) *Analytical Philosophy. Second Series*, Oxford, Blackwel, pp. 1-19; D. M. Taylor (1966) The incommunicability of content, *Mind* 75, 527-541; S. Shoemaker (1969) Time without change, *Journal of Philosophy* 66, 363-381; W. Lycan (1973) Inverted spectrum, *Ratio* 15, 315-319.

[3] Una interpretación de este parágrafo en el contexto del análisis wittgensteiniano de las sensaciones se encuentra en M. Díaz (2010) "Sobre el lenguaje y las sensaciones en la filosofía de Wittgenstein".

CAPÍTULO III

Auto-referencia y el problema de la conciencia en la vida y obra de Francisco Varela

Roberto Ferrari y Ricardo Pulido

Introducción

La obra de Francisco Varela se caracteriza por tres planos de desarrollo completamente integrados entre sí: el científico, el filosófico y el existencial. En el centro de los tres está siempre el fenómeno de la vida y la consciencia, investigadas con métodos y marcos teóricos propios de la cibernética, las ciencias biológicas, las neurociencias, la fenomenología, la introspección y a través de su propia experiencia de la práctica meditativa, la enfermedad y la muerte. En estos tres planos de investigación, Varela se movió desarrollando dos modalidades paralelas o estereoscópicas de conocimiento: el conocimiento transitivo, que se dirige hacia el objeto y trata de describirlo y conocerlo de manera objetiva, y el conocimiento vivido, inmediato, que pone en primer plano más bien la relación que se establece con el objeto antes que el objeto –o el sujeto– en sí mismo. En las siguientes páginas revisaremos el trabajo de Varela a la luz de una idea matriz que atravesó toda su vida y pensamiento: la auto-referencia. Veremos cómo esta idea se despliega, adapta y profundiza a los diferentes campos del saber en que se interesó: las ciencias biológicas, la inmunología, las ciencias cognitivas y en particular en el estudio de la consciencia y el significado de la existencia humana.

Ciencias biológicas y autopoiesis

La pregunta inicial que impulsó el trabajo de Francisco Varela desde sus

inicios como biólogo hasta sus posteriores desarrollos en el campo de la epistemología, ciencias cognitivas y neurociencias fue la pregunta por el estatuto de la conciencia en el universo (M. Bitbol, comunicación personal), entendiendo por consciencia, como veremos más adelante, no una función cognitiva, una estructura nerviosa o un contenido de consciencia (p. ej.: una representación consciente) sino el hecho mismo de la experiencia vivida subjetivamente. A la luz de este interés general inció su recorrido científico orientando su investigación hacia la biología de la percepción y la cognición, partiendo por el estudio del sistema visual de la rana junto a Humberto Maturana.

A partir de esta investigación, Maturana y Varela maduraron durante la década del 70 una de las ideas por la cual han sido mundialmente reconocidos: la teoría de la autopoiesis. En esta se muestra por primera vez que la auto-referencia no es una de las tantas funciones de la vida, sino por el contrario, su característica más esencial. La pregunta desde la que parten es la siguiente: ¿Qué es aquello que puede dar cuenta de la naturaleza de cualquier sistema vivo de la manera más simple y general? Y la van a responder argumentando que es precisamente la capacidad de referirse, antes que a los estímulos externos, a su propia forma y dinámica interna. En términos técnicos: un sistema vivo se caracteriza por estar cerrado operacional y organizativamente. Esta clausura operacional –así la denominan los autores– se puede ilustrar con la imagen de la causalidad circular, donde los procesos metabólicos y las estructuras biológicas se determinan recíprocamente. Esta no es una idea fácil de pensar dado que la tendencia más común en la biología es considerar que los procesos están contenidos en las estructuras (Maturana y Varela, 1985).

De este modo, los biólogos chilenos intentan pensar lo viviente no solo en términos de causalidad *top-down* o de-arriba-a-abajo (por medio de la cual las estructuras celulares determinan los procesos metabólicos locales), sino que en paralelo consideran una causalidad *botton-up* o de-abajo-hacia-arriba (donde, al mismo tiempo, la red de procesos metabólicos determina la constitución de las estructuras físicas). Estas dos vías de causación generan un círculo cerrado de interacción recíproca donde, no obstante el cierre operacional, el sistema se mantiene abierto a los influjos ambientales (lo atraviesan señales, materia, energía). Es así como la manera en que un

sistema vivo procesa los estímulos o perturbaciones del ambiente está determinada por sus propias constricciones organizativas internas y no por los acontecimientos externos en sí mismos. Más aún, estas constricciones organizativas constituyen la identidad misma del sistema y por ello, en su devenir, el sistema viviente se esfuerza constantemente por auto-mantenerlas.

Los seres vivos, entonces, son aquellos sistemas muy peculiares que gracias a sus permanentes procesos o ciclos de auto-referencia se auto-producen (*autopoiesis*) y se auto-estabilizan. Para Maturana y Varela esta nueva concepción de lo viviente abría el camino para comprender el paso de la materia orgánica a la vida y de esta última a la cognición. Es así como reemplazan la clásica noción de que lo vivo está operacionalmente dirigido hacia fuera, por la idea de que lo vivo es un sistema autónomo, capaz de producir por sí mismo tanto las estructuras físicas (membranas celulares, moléculas de ADN, proteínas y enzimas) como los procesos metabólicos que estas estructuras regulan y a partir de los cuales, a su vez, se constituyen. Y desde estos ciclos de auto-referencia no solo surge la vida, sino también la cognición, término que indica "conocimiento" espontáneo generado a partir de los procesos de auto-mantención. Es decir, vida y cognición son eventos coincidentes, no se da uno sin el otro (Maturana y Varela, 1985).

Este cambio de perspectiva cobra relevancia en cuanto modifica la concepción tradicional de que la cognición es un proceso secundario a la vida, un derivado que surge a partir de la complejidad de las estructuras biológicas. Desde esta perspectiva, la vida/cognición, esta unidad indisoluble, no está distribuida en una serie de ciclos y procesos, sino que por el contrario está sintetizada en una forma de self, en una identidad que le da coherencia y unidad a todo organismo. En palabras de Varela: *"The domain of deformation that the system can be submitted to without loosing its identity (i.e. and still maintains its organization) is the domain of transformation where it exists as a unity*[1]*"* (Varela, en Rudrauf *et al.*, 2003, p. 31). Es decir, mientras el organismo o sistema vivo está auto-organizado y no está sometido a perturbaciones excesivas, conserva una identidad estable en el tiempo y de este modo es capaz de tener un conocimiento de sí y de su ambiente, no obstante dicho conocimiento sea un conocimiento directo de los ciclos que lo conforman y no una representación de ellos. De

hecho, lo que caracteriza la muerte de un sistema vivo es precisamente la pérdida de esta identidad fluida y de la actividad cognitiva constituida desde el interior.

Es una visión de la vida alternativa a las ideas funcionalistas, aún en boga, que representan a lo vivo como una serie de estructuras físicas y funciones (mecanismos nerviosos, homeostáticos, metabólicos, reproductivos, etc.) determinados a partir del ambiente externo y cuya utilidad, la de las funciones, es procesar estímulos provenientes del medio ambiente y producir respuestas conductuales en y para el medio ambiente. Pero desde el punto de vista de la autopoiesis, donde la vida es entendida como una red unitaria y auto-organizada, esta perspectiva funcionalista es más bien una estrategia del conocimiento objetivo que busca aislar y representar partes y mecanismos separados para así poder observarlos y eventualmente manipularlos.

Precisamente la dinámica de auto-referencia –idea que tiene su origen en el concepto de retroalimentación negativa o *feedback* de auto-estabilización de los sistemas cibernéticos teorizados en los años 50 y 60– va a constituir la piedra angular para la generación de identidad y capacidad cognitiva del organismo, es decir, la quintaescencia de la característica fundamental y más básica de los sistemas vivos, sean estos células bacterianas o lagartijas. Es una descripción simple y elegante, parsimoniosa, en el sentido que apela solo a este principio autodefinido desde el cual emergen elementos que se producen el uno al otro: los seres vivos se transforman a sí mismos de manera tal que su propia organización es el fruto de su misma operatividad (Varela, 1979).

De esta manera, la auto-referencia es concebida como un principio que produce sistemas vivos y conocimiento en su sentido más amplio. Sin embargo, no se trata de una sustancia inicial o de un agente productivo originario; este principio auto-referencial, tanto como sus productos, está desprovisto de sustancia autónoma y separada, pues existe solo como relación entre los elementos que ella misma genera. En cuanto observadores, no nos es posible aferrar ni estructuras, ni procesos, ni relación entre ellos: la auto-referencia se revela más bien como un principio infundado que no tiene un punto de partida en la materia (“estructura”) o en la información (“procesos”), sino que existe solo en un “girar y dar vueltas” que produce identidades vivientes junto a sus respectivos mundos, los

cuales, no obstante inestables, están constantemente reemergiendo.

Es un planteamiento abstruso y particularmente difícil de divulgar pues no encaja bien en los márgenes paradigmáticos de los actuales conocimientos científicos en donde usualmente se intentan determinar principios generales y leyes naturales a partir de los cuales desciendan y sean predecibles los fenómenos particulares (Bitbol, 2002). En contraposición a esto, con la teoría autopoyética comparecen los primeros rasgos de un nuevo paradigma del conocimiento radicalmente anti-fundamentalista: no se convocan ni planos metafísicos ni leyes naturales, solo la recursividad de los ciclos de auto-referencia desprovistos de sustancia propia y la total incertidumbre de los resultados de esta pulsión generativa.

Inmunología, neurología y enacción

Los trabajos revisados en el apartado anterior permiten a Varela trazar un modelo de pensamiento –el de los ciclos recursivos de auto-referencia– que extenderá al estudio de las redes celulares, la inmunología y de la cognición humana. Por ejemplo, a partir de este modelo, en los años 80 repensó los sistemas nervioso e inmunológico: ambos sistemas serían redes complejas de células asociadas entre sí por medio de señales específicas (químicas o eléctricas) en “asambleas” temporales que producen una permanente retroalimentación (*feedbacks*) de información y auto-referencia y que generan así auto-oscilaciones de señales capaces de absorber las perturbaciones internas y externas (Varela, 1995). Una peculiar nube de ritmos y oscilaciones que conserva su propia organización y forma y que constituye la identidad molecular y cognitiva de dicha red de células, tal como en una colonia de insectos sociales donde una compleja red de relaciones, contactos, señales e intercambios de alimento constituyen una organización extremadamente sofisticada y dotada de una identidad propia y unitaria, es decir, una verdadera “mente colectiva” (Ferrari, Pulido y Ferri, 2006). Dado que esta identidad es capaz de auto-mantenerse absorbiendo perturbaciones incluso intensas, el observador, que siempre busca reificar elementos que justifican la coherencia del sistema, la percibe como estable y definida.

Con tales trabajos, Varela comienza a dar forma a su teoría sobre la experiencia consciente y que lo llevará más adelante a plantear la neurofenomenología como método de indagación. No obstante, hasta aquí la experiencia consciente está pensada como la experiencia del "observador" que describe el modelo y construye la teoría, es decir, está contenida en cuanto condición de posibilidad de toda producción de conocimiento, pero aún no es propiamente tal el tema de sus trabajos científicos. Será en el contexto del diálogo con la filosofía budista y sobre todo incorporando algunas prácticas de indagación en primera persona (p. ej.: meditación shamatta) como fuente válida de producción de conocimiento para las ciencias congitivas, que Varela se abocará a la descripción y estudio de la cognición humana y la experiencia consciente. Su principal desafío va a ser desarrollar un modelo teórico, fundado en la evidencia, que le dé cabida a la experiencia vivida subjetivamente dentro de la cognición humana. Para ello, no podrá buscar la mente separando el cerebro de las redes de relaciones que establece con el cuerpo y el medio ambiente. Tampoco podrá subdividir el cerebro en distritos y funciones definidos en base a daños selectivos. Para Varela, "mente" es algo mucho más básico y originario que estos mecanismos, es la presencia para sí mismo que encuentra una de sus condiciones de posibilidad en las células singulares, en los sistemas celulares o en los órganos como el cerebro. Una presencia que se mantiene aun cuando las funciones cognitivas se interrumpen. En palabras de Francisco: *"We now suggest that the process crucial for consciousness cut across the brain-body-world divisions, rather than be limited to neural events in the head*[2]*"* (Thompson y Varela, 2001, p. 422).

Al inicio de los años 90 este desafío lo llevó finalmente a plantear la teoría de la enacción, donde introduce precisamente la experiencia vivida dentro del campo de la biología como un elemento que efectivamente está causado por el nivel celular y físico, pero que a su vez causa, constituye y modula dichos sustratos biológicos. En esta teoría –desarrollada en *The Embodied Mind*, escrito en colaboración con el filósofo Evan Thompson y la psicóloga Eleanor Rosh– la co-determinación interna entre células y redes celulares se extiende y se asocia con el comportamiento sensorio-motor, inaugurando así una co-determinación de segundo nivel donde la percepción está causada por el moverse y el actuar y, viceversa, el actuar

está causado por la percepción, y esta causación recíproca ocurre en millones de ciclos recurrentes que son la más simple y profunda unidad de experiencia de lo viviente (Varela, Thompson y Rosch, 1991). Estos ciclos de percepción-acción, que para Varela son ya las primeras vivencias fenomenológicas describibles, modulan las dinámicas de las estructuras biológicas, pero a su vez se ven limitados y enriquecidos por estas últimas, por ejemplo a través de específicas posibilidades de percepción o movimiento.

Hasta aquí tenemos entonces dos niveles de co-determinación (o ciclos de auto-referencia). El primero, está en un dominio biológico de descricpción transitiva (células, metabolismo, etc.), mientras que el segundo es ya un dominio fenomenológico de la experiencia vivida (percepción senso-motora). Por último, un tercer nivel de co-determinación completa el modelo enactivo. Este nuevo ciclo de auto-referencia se establece entre la identidad cognitiva del organismo (que emerge a partir de los ciclos senso-motores) y el "mundo", pero no con el ambiente "objetivo", sino con la experiencia del ambiente o del entorno (ale. *umwelt*) que cada ser vivo construye y con el cual interactúa intercambiando energía y conocimiento.

Gracias a la recursividad de los ciclos de co-producción se crean entidades que incluso cuando parecen estables y reconocibles, no tienen una sustancialidad intrínseca e inmutable, sino que por el contrario, fluctúan y solo pueden ser asiladas aproximativamente. De hecho, si bien podemos reconocerlos, no podemos determinar de manera fija los extremos de los distintos niveles en que se dan los ciclos de auto-referencia pues cada uno de estos polos existe solo en relación con el otro (célula y red de células; sentir y actuar; organismo/identidad y ambiente/mundo).

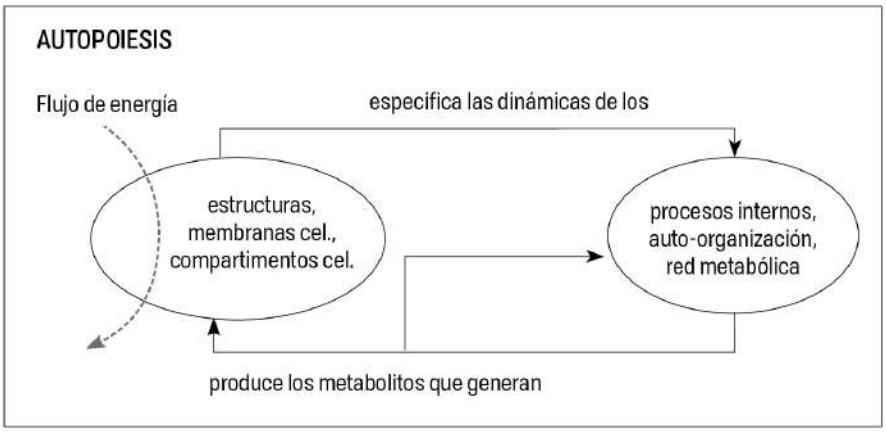
Este tipo de pensamiento tiene sus raíces no solo en la noción de emergencia y en la cibernética, sino también en pensamientos mucho más antiguos como el del filósofo budista Nagarjuna. En su obra *Madhyamaka Karica*, Nagarjuna niega la posibilidad de que los entes tengan alguna "sustancia propia" (sscr. *svabhava*), es decir niega que puedan poseer una esencia característica y definitiva que hace que una cosa sea precisamente esa cosa y de este modo, se aleja tanto del realismo (hay un mundo exterior que crea organismos y al cual estos se adaptan) como del idealismo (solo existe el sujeto viviente que proyecta un mundo). En términos científicos, investigando la naturaleza del mundo y del sujeto que lo habita, Varela

afirma que no hay sustancia ni en el uno (mundo) ni en el otro (sujeto/identidad). Ambos existen como una escenificación o "puesta en escena" (ing. *enaction*) que surge a partir de los ciclos recurrentes de acción y percepción encarnados (ing. *embodied*) y situados en un contexto. Este carácter de la experiencia de *embodiment* o de estar-situada se expresa fielmente en el lenguaje heideggeriano con el término ser-ahí (ale. *da-sein*), o ser-allá... allá donde la experiencia tiene lugar (Varela, 1991). Y al mismo tiempo, posee también el carácter de la constricción, una prisión radical: "*We are taken in a cognitive system, from which we can't neither escape nor choose where it begins or how it works*"[3] (Varela en Rudrauf *et al.*, 2003, p. 40).

Una pregunta que se abre es acerca de la compatibilidad de esta perspectiva con el punto de vista histórico y consolidado de la teoría de la evolución. De hecho, es precisamente gracias a la falta de sustancia propia que los seres vivos se pueden transformar y adaptar con gran eficiencia. La evolución no es solo la evolución de un organismo objetivo, inmerso en un ambiente y un tiempo histórico, sino que está ocurriendo, actuándose a cada instante, como transformación de la auto-organización del organismo que selecciona y lleva hacia su interior la relación con el ambiente y en tal forma va construyendo su propio mundo "suficientemente" adecuado (Varela, Thompson y Rosh, 1991). En otras palabras, es a raíz de esta capacidad de transformación, fundada en la insustancialidad de su estructura y proceso, que el ser vivo es capaz de absorber las perturbaciones del medio ambiente. Cada adaptación deja huellas que luego interpretamos como evolución histórica. En última instancia se trata de una co-evolución, pero no entre un sujeto y el mundo, sino entre las experiencias de acciones y percepciones del organismo mismo. Sin embargo no se trata de las experiencias de un sujeto que actúa y percibe, pues esto supondría una identidad (o *self*) perceptivo y agente, cuando en cambio, la identidad surge a partir de la co-determinación de las acciones y percepciones mismas.

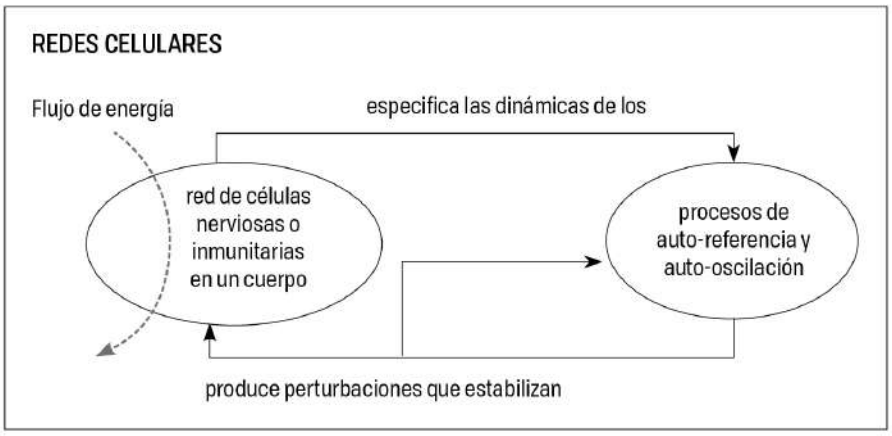
La perspectiva enactiva es un desarrollo sucesivo fundado en la misma estructura auto-referencial de la precedente teoría autopoyética. Tal como muestra el Esquema 1, en la autopoiesis se da una co-determinación entre estructura y auto-organización de los procesos metabólicos de la célula, una auto-referencia operacionalmente cerrada y a la vez abierta a la circulación de energía que atraviesa el sistema.

Esquema 1
Co-determinación autopoyética



Posteriormente, esta misma estructura se amplía al nivel de las redes celulares, donde la co-determinación se da entre las conexiones físicas y los procesos de auto-resonancia de las asambleas celulares.

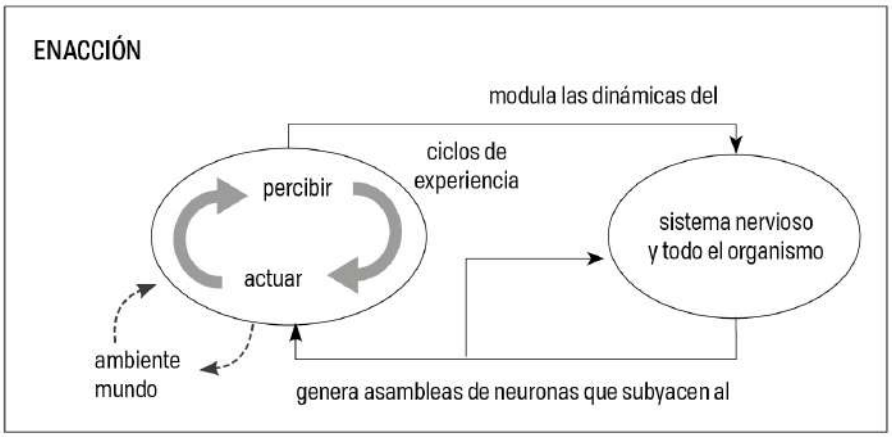
Esquema 2
Co-determinación de las redes celulares



Finalmente, en la enacción, la co-determinación ocurre entre todo el organismo y su modo de moverse y percibir un mundo. Este comportamiento auto-referido representa la irrupción de la experiencia vivida del organismo ahí donde antes solo se postulaban procesos metabólicos, auto-organizativos o auto-oscilatorios. Es la experiencia la que co-produce realidades compuestas y distribuidas tales como el cuerpo,

la identidad, el mundo.

Esquema 3
Co-determinación enactiva



La enacción es un desarrollo de las teorías de la emergencia, en el que la cognición humana no surge "desde abajo" (del nivel celular o físico), sino que es una "puesta en escena" tanto "desde arriba" (del nivel global de la experiencia vivida) como "desde abajo". Varela se daba cuenta que tres importantes campos del conocimiento podían revitalizarse gracias a su nuevo enfoque de biología teórica:

1. Desde una perspectiva epistemológica, las auto-referencias tienen una doble implicación: por una parte muestran que todo conocimiento, así como la adaptación y flexibilidad de la vida, se fundan en esta circularidad generativa; por otra, evidencian que la circularidad impone un límite para el conocimiento objetivo, en cuanto disuelve toda posibilidad de fundamento y realidad de los entes aislados. Se desvanece así la esperanza de explicar el mundo, la vida y la mente como derivados de una sustancia inicial o de definirlos por medio de representaciones, puesto que toda co-determinación emerge sin un principio y sin un fin. Es posible reconocer estos fenómenos, incluso describir sus estructuras, pero si se intenta explicarlos (por ejemplo: "el espíritu o la materia preceden la consciencia") o definirlos e historizarlos (por ejemplo: "la vida es el resultado de una creación o de una evolución biológica"), estas mismas explicaciones recaen dentro

de la circularidad generativa: ¿quién lo está sabiendo, el espíritu o la materia? ¿cómo lo sabe? (Bertossa *et al*., 2004; Bertossa y Ferrari, 2005; Bitbol, 2001).

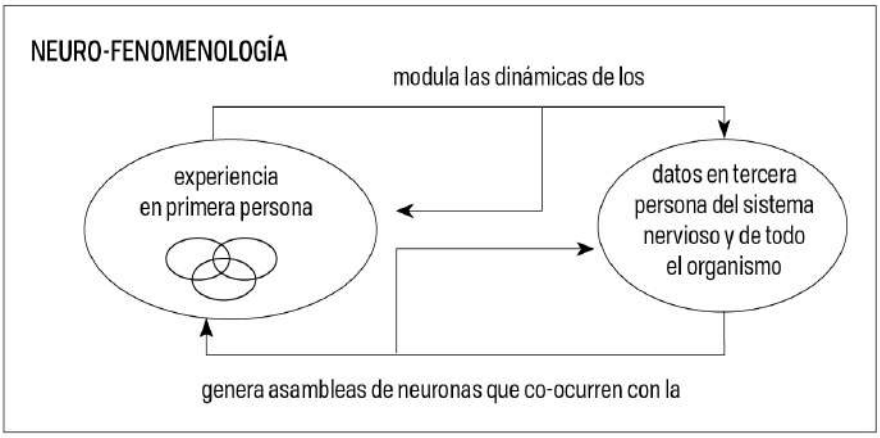
2. Desde el punto de vista científico, la dimensión de la experiencia finalmente ingresa en las investigaciones de laboratorio de las ciencias cognitivas. Varela se negaba a postular que la experiencia existiese en un tiempo o en un lugar definidos, que tuviese causas físicas directas y que pudiese ser aferrada. Para él, se trata más bien de un evento transitorio y de naturaleza distribuida producido, precisamente, por los ciclos de auto-referencia entre cuerpo y cerebro, sentir y actuar, sujeto y mundo. De este modo, como veremos más adelante, logró impulsar un campo de investigaciones vanguardistas sobre la integración temporal de los momentos de conciencia inestables que se vinculan con la sincronización de asambleas de células nerviosas sintonizadas a la distancia.
3. Por último, en términos de la experiencia directa, las realidades sin fundamento y sin estabilidad que surgen de los ciclos auto-referenciales, no pueden ser negadas. Para comprobarlo es suficiente examinar en la vivencia misma la identidad individual "yo", que aún desprovista de un *self* sustancial y estable, constantemente se afirma y en ningún modo se deja negar: "*This global level of me as an individual appearing in our mind-related interactions is 'a mode of existence of which you cannot say it doesn't exist ('Francisco doesn't exist')*"[4] (Varela en Rudrauf *et al*., 2003). ¿Dónde se encuentra entonces este yo? Desde el punto de vista teórico incluso el *self*-identidad es una construcción virtual: no lo podemos hallar. Pero desde un punto de vista fenomenológico, Varela afirmaba que aquello que impulsa el surgimiento de este nivel global de la identidad o del yo es el hecho que los organismos son existencias que siempre están siendo empujadas por la falta de significado que ellos mismos generan afirmando su presencia (Varela, 1991). En la experiencia del "yo", como en la de otras realidades "puestas en escena" por los ciclos de auto-referencia, conviven paradojalmente la vivencia de la innegabilidad de sí mismo con la viviencia de la propia ausencia de sustancia y significado. Es una pulsión que no se puede eliminar y que

se manifiesta como cognición: "*La cognizione è azione riguardo a ciò che manca, riempimento delle manchevolezze dalla prospettiva di un sé cognitivo*"[5] (Varela, 1992, p. 67).

Neuro-fenomenología

¿Cómo investigar la mente desde una aproximación científica sin sacrificar la experiencia subjetiva en primera persona? En la última parte de su vida, Varela se esforzó por desarrollar una nueva metodología que fuera capaz de estudiar la experiencia consciente en el laboratorio. La búsqueda de una naturalización no reduccionista de la mente lo impulsó a convocar, en París, a muchos estudiantes e investigadores de neurociencias interesados en la dimensión de la experiencia y, junto a ellos, introdujo el análisis fenomenológico al interior de las ciencias cognitivas. El proyecto de la neuro-fenomenología consiste precisamente en aparear –nuevamente en una co-producción generativa de conocimiento– los datos científicos objetivos en tercera persona con datos igualmente detallados y confiables producidos examinando la experiencia en primera persona (Varela, 1996). Este examen se basa en la capacidad auto-referencial de nuestro flujo mental. Tal como muestra el Esquema 4, por una parte tenemos el polo de la experiencia del sujeto al cual solo podemos acceder por medio de datos en primera persona (p. ej.: reporte de experiencias subjetivas) y por otra, las estructuras y funciones del sistema nervioso del organismo al cual accedemos con datos en tercera persona (p. ej.: electroencéfalograma).

Esquema 4
Co-determinación de la neuro-fenomenología



El supuesto de la neurofenomenología es que ambos polos mantienen una relación de dependencia recíproca: el primero modula las dinámicas del segundo y el segundo genera asambleas neuronales que emergen en co-ocurrencia con las experiencias en primera perosna.

Este enfoque de la neuro-fenomenología permite disolver el clásico "*hard problem*" en el estudio de la naturaleza de la experiencia consciente en la medida que prescinde del hábito "esencialista" de definir la conciencia (por ejemplo como una emergencia de la materia, información, espíritu, etc.). El *hard problem* se crea precisamente por medio de un "camino ontológico" de reificación del fenómeno de la consciencia: para comprender la experiencia consciente se busca definir "qué es" la consciencia (es decir su ontología) para luego estudiarla y describir cómo emerge la experiencia. Pero precisamente al definirla se crea el problema, pues se fija la consciencia en una esencia (p. ej.: materia, energía, información, espíritu, etc.) y entonces no se logra explicar cómo desde esta esencia nace la experiencia. Varela, en lugar de proponer nuevas soluciones esencialistas, echa por tierra el planteamiento mismo del problema, proponiendo suspender toda definición y reemplazar esta "vía ontológica" por un "camino metodológico" que se mantenga fiel al fenómeno mismo y que permita incluir en el campo de la ciencia la co-determinación entre informes de primera persona y reportes en tercera persona, entre datos experimentados por un sujeto situado y mediciones experimentalmente controladas.

Una revisión de sus trabajos científicos más significativos nos permitirá aclarar este nuevo enfoque metodológico. Varela y su equipo se preguntaban dónde buscar las huellas de un acto singular de percepción y

reconocimiento (ing. *perception*). Partieron intentando aclarar fenomenológicamente ¿cómo surge el instante del reconocimiento? Recogiendo así las descripciones en primera persona asociadas a la experiencia perceptiva y analizando las estructuras invariantes que cruzan las distintas vivencias individuales, lograron determinar que el reconocimiento aparece como una integración rápida y transitoria de diversos contenidos de conciencia en una escena y que dicha escena representa una unidad temporal, una percepción/comprensión que no se puede comprimir ni expandir en el tiempo y que está entrelazada con en el sentir del cuerpo (Varela, 1995). Esta descripción los llevó a plantearse la hipótesis de que los procesos cerebrales (datos en tercera persona) que acompañan esta experiencia podrían deberse a la integración transitoria de los tiempos de descarga de las neuronas, incluso en zonas del cerebro muy distantes entre sí. Esta hipótesis, de hecho, permitía explicar tanto la unidad de tiempo y espacio que se requiere para identificar el instante singular de reconocimiento, como la transitoriedad del mismo, condición necesaria para dar cabida a nuevos momentos de conciencia ("presente cognitivo").

Esta hipótesis construida a partir de los informes fenomenológicos y del estado del arte de las neurociencias cognitivas, fue testeada experimentalmente por Varela y sus colaboradores (Rodríguez *et al.*, 1999). Así descubrieron que cuando un sujeto reconoce una forma ambigua (como "caras lunares" presentadas al revés), no sucede nada en los instantes apenas iniciales, pero luego, más menos 250 ms después de la presentación de la imagen, se evidencian "asambleas de células" distantes entre sí que sincronizan la frecuencia de sus descargas nerviosas. A través de mediciones de EEG, notaron una sincronización de fase a gran escala en el instante del reconocimiento, especialmente en la banda de las ondas gama (descargas con frecuencia de 30 a 80 hertz) seguida tanto por una dispersión de fase (aproximadamente después de 400 ms) en la que las áreas anteriormente conectadas se desincronizan dando lugar a un nuevo momento cognitivo, como por una nueva sincronización entre asambleas de células distantes (unos 600-700 ms después) que se corresponde con la respuesta motora del sujeto (la presión del botón para señalar que ha reconocido la imagen). Por el contrario, cuando un sujeto no puede reconocer la imagen, no se producen sincronías o desincronías de asambleas

celulares, evidenciando de este modo que la experiencia de responder, sea esta cognitiva o motora, está siempre coordinada con sincronizaciones de fase transitorias y muy rápidas a las que Varela llamó “las sombras de la percepción”.

Otra estudio importante conducido por Lutz y colaboradores (2002) exploró las diferentes actitudes de personas que participan en el reconocimiento de figuras, particularmente estereogramas (figuras de las que tras una observación cuidadosa puede surgir, repentinamente, una imagen en tres dimensiones). Justo antes del experimento, los sujetos realizaban una entrevista de auto-descripción fenomenológica que permitía distinguirlos en dos grupos: aquellos que se percibían a sí mismos como preparados y atentos para la tarea y aquellos que se sentían distraídos. Los datos neuronales fueron distribuidos siguiendo la división de los dos grupos, apareando de este modo el dato fenomenológico con el neurológico. Y efectivamente las respuestas son muy diferentes: los sujetos “preparados para la respuesta” mostraron sincronizaciones de fase incluso algunos instantes antes de la presentación del estímulo, de manera que las asambleas neuronales del reconocimiento emergían rápidamente (en promedio 273 ms antes de la presentación de la imagen) y permanecían sincronizadas por mayor tiempo. Esto coincide también con la vivencia: el sujeto experimental vive el instante del reconocimiento de la imagen con una sensación de estabilidad y continuidad respecto de los momentos anteriores. En contraste, los sujetos “no preparados y distraídos” no presentan momentos cognitivos unificados antes de la presentación del estímulo y luego de ella solo se activan patrones débiles de sincronización que se disuelven inmediatamente (son como intentos de reconocimiento que no alcanzan a ser exitosos). Posteriormente, alrededor de 300 ms después de la presentación de la imagen, registran una intensa desincronización entre las zonas frontal y occipital, lo que representa una vuelta a un estado neutro de base (ing. *reset*) y una oportunidad de reorganización completa. Por último, en estos sujetos emerge “lentamente” (en promedio alrededor de 520 ms desde la presentación del estímulo) una asamblea de células sincronizadas en el área frontal en paralelo al momento del reconocimiento de la imagen en 3D. Y respecto de la vivencia del proceso, los sujetos “no preparados y distraídos” la describen con un sentimiento de sorpresa, es decir como una clara ruptura respecto de los instantes anteriores.

Finalmente, otro ejemplo de neuro-fenomenología que tiene insospechados ribetes de aplicación terapéutica, es un estudio sobre las regularidades que se manifiestan en las crisis epilépticas tanto en el nivel de las experiencias en primera persona como en el de los registros cerebrales. Varela y sus colaboradores evidenciaron que gracias a las relaciones recíprocas que se establecen entre la activación neuronal objetiva y la experiencia subjetiva –considerada ya no como un mero epifenómeno de una enfermedad fisiológica sino como un nivel susceptible de ser descrito rigurosamente y que tiene efectos causales sobre el plano neuronal– los pacientes epilépticos, reconociendo sutiles señales que se manifiestan en el plano subjetivo, son capaces de aprender a detectar los pródromos de la crisis varias horas antes de que ocurra. De esta manera, el paciente no solo puede ponerse a salvo para controlar la violencia del ataque, además puede aprender a estabilizar ciertos estados mentales específicos que le permitan inhibir la crisis epiléptica (Le Van Quyen y Petimengin, 2002).

Si bien estos estudios de neuro-fenomenología permitieron explorar algunas propiedades de la conciencia –(a) la creación de un momento cognitivo unificado y su transición al siguiente; (b) la actividad espontánea de auto-sincronización del sistema nervioso gracias a la atención y el entrenamiento; y (c) las terapias cognitivas de trastornos que clásicamente han sido tratados dese un punto de vista exclusivamente nervioso– una mirada más atenta a ellos revela que los resultados a los que se llega con esta metodología experimental no permiten comprender la consciencia en sí misma, más allá o en independencia de sus atributos, propiedades y funciones. En otras palabras, estos estudios todavía no arrojan luces sobre qué es ese “darse cuenta”, ese instante situado y sentido en el cuerpo (*embodied*) que constituye la toma de consciencia, el conocer, el saber. Consciente de esto, en la última etapa de su vida Varela intentó impulsar su programa científico-filosófico a niveles cada vez más sutiles y refinados de la descripción fenomenológica de la experiencia consciente. Concretamente, su interés se concentró en la comprensión del instante mismo del encontrarse siendo-ahí (*being-there*) en el momento en que algo concreto y específico se muestra (Varela, 1995).

Tres gestos para volverse consciente

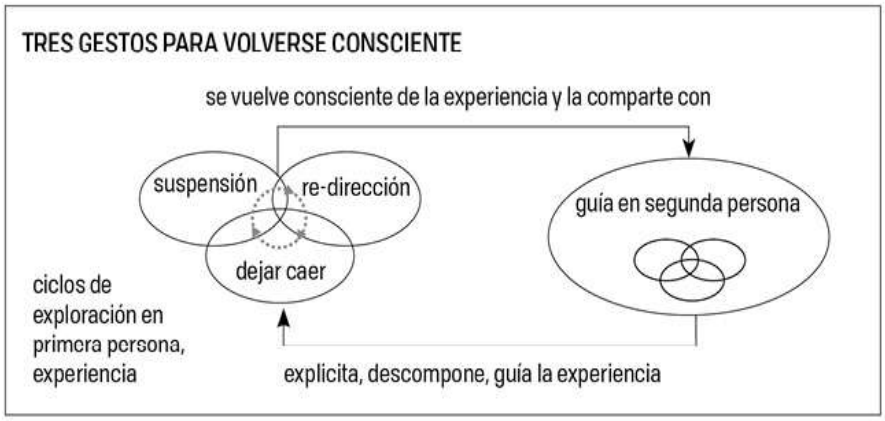
Francisco Varela alcanzó a plantear esta cuestión en sus últimos años. Fue el desarrollo natural del programa neuro-fenomenológico y, al mismo tiempo, la promesa de un salto decisivo hacia adelante. Comprende que para experimentar y poder explorar estos momentos, repentinos e inestables, en los que nos damos cuenta y nos volvemos conscientes se necesitaba introducir en la reflexión metodológica tres aspectos relacionados entre sí: 1) abandonar la visión ingenua del método introspectivo que confía plenamente en los reportes de la experiencia del sujeto; 2) perfeccionar métodos de acceso a la experiencia basados en el entrenamiento de destrezas y habilidades descriptivas que puedan dar cuenta de experiencias como la auto-referencia, la co-emergencia de los fenómenos y la falta de fundamento del conocimiento; y 3) producir de este modo datos de experiencia en primera persona que posean una confiabilidad y exactitud equivalente a la de los datos en tercera persona, permitiéndoles de este modo salir del ámbito de la "experiencia privada" y ser valorados en cuanto conocimiento compartido inter-subjetivo.

El segundo de estos puntos reviste un rol crucial en cuanto demanda el desarrollo y perfeccionamiento de una praxis de acceso a la experiencia. Si bien la fenomenología nace en Occidente precisamente como un método para acceder y explorar la experiencia misma, Varela (quien desde muy temprana edad cultivó la lectura de Husserl, Merleau-Ponty y otros) sostenía que nuestros fenomenólogos más notables, siendo maestros naturales y verdaderos *enfant prodige* de la "ciencia de la experiencia", produjeron descripciones precisas, pero sin una sola línea de instrucciones prácticas acerca de cómo replicar las experiencias que se encuentran a la base de tales descripciones. Para remediar esta falta, exploró en el patrimonio de conocimiento de las tradiciones contemplativas de Oriente, principalmente de las dos más antiguas y sistematizadas como son el hinduismo y el budismo. Al mismo tiempo, destacó el papel de un guía (al que denominó "segunda persona", o sea un "tú" que sirve como interfaz entre la primera y la tercera persona) que habiendo desarrollado una maestría específica producto de un entrenamiento personal, puede ayudar al neófito a deconstruir su experiencia, explicitar los elementos implícitos de ella y llevar y sostener la atención en la fuente de la experiencia misma. Tal como hacen los expertos de música, deporte o arte cuando buscan transmitir su *know-how*, el guía o "segunda persona" instaura una relación empática

con el sujeto que le permite reconocer, re-experimentando en su propio cuerpo, aquello que el alumno va sintiendo y vivenciando, pues él mismo ya lo ha vivido y sentido innumerables veces.

La importancia de una guía en segunda persona demuestra que este acceso a la experiencia tiene más que ver con una forma de entrenamiento concreto antes que con un punto de vista filosófico o metafísico que debe ser adoptado. Como podemos notar en el Esquema 5, Varela junto a Depraz y Vermersh, distinguieron tres disposiciones, gestos o momentos en el proceso del "darse cuenta" (Depraz *et al.*, 2003).

Esquema 5
Co-determinación de los "tres gestos para volverse conscientes"



El primer gesto es la suspensión, la que se haya presente tanto en la fenomenología como en la tradición budista. Constituye un paso esencial para agudizar la atención y despejarla de juicios, hábitos, esquemas y concepciones para así enlentecer el flujo mental, crear espacio y relacionarse con lo que ocurre bajo la modalidad de lo no-sabido. En esta dimensión es necesario permanecer sin hacer nada (ni siquiera describir las esencias como se hace en la fenomenología), lo que lejos de ser algo sencillo y fácil de lograr, exige desarrollar la capacidad de escuchar y sentir las sensaciones en un estado de plena presencia mental, atención plena o *mindfulness* y aprender a soportar la experiencia del "no-pasa-nada". En este primer paso el cuerpo juega un papel central: gracias a la regulación de las energías psicofísicas, la inmovilidad y el equilibrio centralizado de la postura corporal, se crea una nueva disposición del cuerpo que "contagia"

la mente y le permite cambiar de estado (Bertossa y Ferrari, 2006).

El segundo gesto es el de la re-dirección, presente tanto en el introspeccionismo de Occidente como en las tradiciones hindú y budista Zoo-Chen. Varela entendía este paso como un mover la atención desde los contenidos de la consciencia hacia la fuente del acto que los capta (Varela, 1996). Aquí, una vez más, aparece el principio de la auto-referencia, pero por primera vez no en cuanto postulado teórico, sino como vivencia del acto consciente que se repliega sobre sí. Como hemos visto, una característica de este auto-referirse es la imposibilidad de separar los dos polos de la co-determinación: en este caso ambos polos, la "fuente" y el "contenidos de conciencia", poseen una naturaleza in-sustancial y transitoria, no pueden ser aprehendidos ni objetivados y no puede darse uno sin la presencia concomitante del otro. La experiencia se lleva en forma precisa, progresiva y paciente a un repliegue o auto-referencia vivida cada vez mayor, hasta el punto en que la experiencia consciente misma se impone como una auto-evidencia.

El tercer gesto es típico del budismo. Se trata de un "dejar ir" o "dejar caer", una actitud totalmente opuesta a los hábitos de apego o aferramiento. A menudo se le asocia con una des-identificación de las representaciones, modelos e ideas. Es un paso esencial para escapar de la trampa en que nos dejan todos aquellos intentos por definir la consciencia y que nos llevan a querer aferrarla objetivamente en lugar de simplemente serla. En otras palabras, corresponde a un coincidir con la mirada misma que busca (Bertossa y Ferrari, 2006), pero sin que el gesto de "dejar caer" se desapegue de la motivación a investigar e indagar. Este último gesto diferencia radicalmente la praxis meditativa con finalidad terapéutica o de bienestar respecto de la que se emprende con una finalidad cognoscitiva.

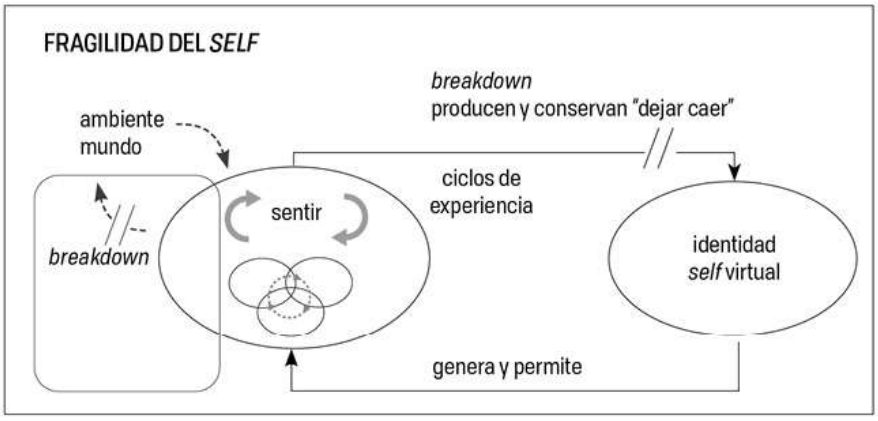
Estos tres gestos se repiten constantemente y se enriquecen entre sí gracias a la guía de la segunda persona: una suspensión más profunda, producida por el dejar ir, permite un repliegue sobre sí cada vez más preciso y un nuevo dejar ir que abre espacios para el darse cuenta y el conocimiento.

La fragilidad existencial

Francisco tenía una clara visión de la dimensión “virtual” de las formas de vida y del *self*. Seguramente porque la sentía como una parte íntima de su propia experiencia, solía referirse a esta dimensión en un modo del todo personal llamándola “fragilidad”. Desde su concepción, nuestra vida está marcada por la inestabilidad, no solo en el tiempo, sino también en la experiencia presente de la identidad central que cada uno de nosotros es, siente y conoce a través del acto de auto-referencia. Esta “fragilidad” de nuestro self tiene dos implicaciones relevantes: por un lado, es una promesa de transformación, toda vez que si fuésemos entidades sólidas y permanentes no sería posible ningún proceso de crecimiento, aprendizaje y renovación; por otro, faltando un self estable, nos hayamos siempre existiendo como experiencias frágiles y fluctuantes, co-emergentes de instante en instante gracias a los esfuerzos vitales de auto-referirnos y auto-conservarnos lo más idénticamente posible a nosotros mismos.

Nuevamente podemos ver que este modelo sigue la estrcutura inical de la autopoiesis. Tal como se aprecia en el Esquema 6 tenemos dos polos que se co-producen.

Esquema 6
Co-determinación de la fragilidad del *self*



El primer polo corresponde a los ciclos auto-referidos de experiencia. Estos son los ciclos de sentir-actuar que se constituyen en su interacción con el mundo/ambiente (ver Esquema 3) y que podemos explorar fenomenológicamente por medio de la circulación de los tres gestos para volverse consciente (representados por los tres círculos intersectados). El

otro polo comprende la experiencia de identidad o de *self* y que en términos varelianos es siempre un emergencia y como tal se caracteriza por su virtualidad (en contraste con una noción esencialista de un yo fijo y permanete) y por ende fragilidad. La co-producción de estos dos polos nos muestra que mientras los ciclos recursivos de sentir-actuar producen y conservan el *self* virtual, este a su vez genera y permite los ciclos de experiencia de sentir-actuar. Es una relación de co-determinación en el dominio fenomenológico que complementa el punto de vista transitivo, donde los ciclos de sentir-actuar se co-producen a partir de estrcuturas biológicas, pero en esta vez lo que aparece en primer plano no es la emergencia, conservación y estabilización de la identidad virtual del *self*, sino por el contrario, su constante "fracaso" por afirmarse definitivamente y de una vez a raíz de los continuos quiebres o *breakdown* en la experiencia de sí-mundo que obligan a reformular no solo la propia identidad sino también la relación con el ambiente/mundo.

Surge entonces una pregunta importante que el mismo Varela enfrentó al final de su vida: ¿Cómo nos damos cuenta, desde dentro del sistema, de esta omnipresente fragilidad? ¿Cómo sabemos de todo esto? Si bien Varela intentó dar respuesta desde muchos ángulos a esta pregunta, nunca esgrimió como argumento la existencia de una consciencia o sustancia eterna. En lugar de ello, se mantuvo fiel a su propia experiencia y se dio cuenta de la fragilidad del yo precisamente en esos momentos de colapso cognitivo (*breakdaown*) en los cuales la mente se des-estructura por un momento y caen la obviedad y la transparencia del mundo (Rudrauf *et al.*, 2003). En Varela, esta descripción de la experiencia no solo se vinculaba con su investigación, tanto desde el punto de vista científico (en el cerebro los *breakdown* que siguen perturbaciones incomprensibles corresponden a desincronizaciones y emergencias de nuevas asambleas de células que las reconocen y absorben) como filosófico-metodológico (el gesto de "dejar caer" resetea las estructuras transitorias y nos des-identifica de ellas), sino que por sobre todo, se vinculaba con su propia vivencia existencial: *"'Letting go' is an interesting gesture, because in fact it's almost like invoking the virtuality of the self, just putting it spontaneously on the table. Usually it's life that makes you let go. You know what I mean, in the extreme cases of sickness or danger, or the disappointment of love, it just forces you into that gesture of letting it be, letting it go. It's interesting that human*

beings do have the capacity to mobilize that capacity all the time"[6] (Varela, en Scharmer, 2000, p. 12).

Sabemos de nuestra fragilidad no por un eventual contraste con una sustancia permanente, sino por las constantes vivencias de quiebres o colapsos de la recognocibilidad y obviedad del mundo. Estos *breakdowns* le dan el impulso a los ciclos de auto-referencia generativa para que reconstruyan rápidamente el mundo y la identidad y al mismo tiempo, constituyen aquello que se muestra cuando –viviendo situaciones límites o practicando el "dejar ir" en la meditación– recorremos hacia atrás la estructura extrovertida de la conciencia, regresando así experiencialmente al inicio de aquel impulso. Lo que muchos podrían considerar una conclusión nihilista –todo es finitud y dolor por las pérdidas– para Francisco Varela era la base misma de la ética y de una buena vida: "*What is it to have a good life? A life of wisdom is to be constantly engaged in that letting go, and letting the virtuality or the fragility of the self manifests itself. When you are with somebody who really has that capacity to a full- blown level, it affects you. When we meet those kinds of people, it's clear, because the whole process is not individual, it's not private, and you enter into that kind of resonance"*[7] (Varela en Scharmer, 2000, p. 12).

La fragilidad del *self*, auto-referida y por ende conscientemente vivida, permite el surgimiento de una ética del saber-hacer, del *know-how* o *savoir-faire*, que reconoce y actúa espontáneamente hacia el bien sin necesidad de fundarse en teorías o leyes morales, como en cambio ocurre con las éticas fundadas en el saber representativo (*know-what*) (Varela, 1992). Es un saber-hacer que surge de la clara percepción de que no existe un self o un yo separado del resto de la existencia y que sería necesario defender y hacer prevalecer por sobre otros yoes.

Es así como para Varela el conocimiento y la ética no son abstracciones sino acciones encarnadas (*embodied*), un concreto estar-situado en el saber y sentir del cuerpo y en el ser-ahí (da-sain), donde co-emergencia y fragilidad, el surgir y el disolverse de nuestra identidad y de nuestro mundo, son indistinguibles y permanentes. Como escribe en su poema de iluminación y en su posterior comentario un gran maestro del Budismo Chan:

Birth and Death.
Day and Night.

Running water, stagnant pool.
Bud and fading flower.
Can I find the point at which they change
From one into the other?
Can my nostrils turn upwards?[8]

Comentario: "*I suddenly desired to urinate. And again, watching my urine stream out, everything stopped. Experience the eternal moment, I saw with my Buddha Eye. Now I knew. Nothing is born and nothing can die. Everything simply is*"[9] (Han Shan, 1368-1643).

Varela insistía que cuando surge la experiencia de la fragilidad se hace necesario, más que nunca, no olvidar la innegable realidad de que somos experiencia consciente, pues precisamente en aquellos momentos en que la existencia se revela esencialmente insustancial e impermanente es cuando el hecho de ser experiencia consciente permanece y se evidencia aún con mayor claridad. ¿Qué sucede en el momento de la muerte? Esta es una pregunta que Varela formuló libremente en el último período de su enfermedad e intentó elaborarla con su habitual amabilidad y apertura: "Para los tibetanos, existe esta corriente de conciencia sutil que es la base para los otros niveles de conciencia y cognición… Se muestra en momentos particulares, como la muerte. Cuando la mente común comienza a desmoronarse es como si este núcleo de conciencia estuviese resplandeciendo. No es tuya, ya no eres Francisco, pero sigue siendo consciencia, como una corriente que persiste a pesar de que tu cuerpo y cerebro estén descompuestos… En este punto, la ciencia ni siquiera puede concebir esta idea… Desde mi punto de vista, debemos suspender el juicio, dejar el problema en suspenso. En mi vida he aprendido que una de las mayores dificultades es tener paciencia y quedarse con la pregunta abierta" (Varela, en Reichle, 2005).

Conclusión

A través de un pensamiento encarnado y profundo que ha resultado ser una vanguardia absoluta en los campos de la biología y la neurociencia, Francisco Varela consiguió crear una nueva manera de hacer investigación científica, en la que el observador siempre se considera como un sujeto

encarnado (*embodyment*) y situado al interior del proceso de observación y no como portador de un punto de vista abstracto y objetivo. Las piedras angulares de su recorrido fueron los ciclos de auto-referencia que incluyen concretamente al observador en cada observación y que son ciclos que presentan una naturaleza generativa (co-emergencia) y al mismo tiempo in-fundada (transitoriedad y des-estructuración).

Gracias a este enfoque, Varela logró un objetivo que quizás aún no hemos comprendido del todo, pero que nos ofrece importantes desarrollos: (a) desactivó la estéril contraposición entre mente y cuerpo y el problema de cómo uno de estos deriva del otro; (b) propuso un nuevo paradigma de estudio de la experiencia consciente desde la biología y la fenomenología que en lugar de fundarse en teorías y representaciones, se basa en la experiencia ahí donde siempre ha estado: en la vivencia de este momento inalienable e indiscutible; y (c) sistematizó un método para examinar la experiencia del momento presente a través de ciclos conscientes de suspensión, auto-referencia y des-identificación.

Si bien en el plano científico y filosófico su legado será de largo plazo, en el plano existencial ya podemos apreciar el valor de su pensamiento: poner constantemente en evidencia la existencia de la experiencia consciente, una presencia segura, indudable, pero frágil, siempre al borde de la disolución y re-emergencia en forma de identidad. ¿Por qué existe? ¿Por qué existe en lugar de nada? Este gap explicativo –no entre mente y cuerpo, sino entre ser y nada– realmente no se puede resolver… para nosotros es una fuente de asombro constante.

Agradecimientos

Agradecemos especialmente a Franco Bertossa y Michel Bitbol por sus invaluables enseñanzas y guías para adentrarnos a la comprensión de la vida y obra de Francisco Varela.

Referencias

Bertossa, F. y Ferrari, R. (2010). *La mirada sin ojo: Experimentos sobre la mente consciente, entre ciencia y meditación.* Santiago: J. C. Sáez Editor.

—— (2006). Meditazione di presenza mentale per le scienze cognitive. Pratica del corpo e metodo in

prima persona. En M. Cappuccio (Ed.), *Neurofenomenologia*, pp. 271-291. Milano: Bruno Mondatori.

Bertossa, F., Ferrari, R. y Besa, M. (2004). Matrici senza uscita. Circolarità della conoscenza e prospettiva buddhista. En M. Cappuccio (Ed.), *Dentro la Matrice*, pp. 107-128. Milano: Alboversorio.

Bitbol, M. (2001). Non representationalist theories of knowledge and Quantum Mechanics. *Nordic journal of philosophy - SATS*, 2, pp. 37-6.

——(2002). Science as if situation mattered. *Phenomenology and the Cognitive Science*, 1(2), pp. 181-224.

Depraz, N., Varela, F. J. y Vermesch, P. (2003). *On becoming aware: a pragmatics of experiencing*. Amsterdam: John Benjamins Publishing Company.

Ferrari, R., Pulido, R. y Ferri, F. (2006). Menti connettive e produzione di mondi negli Insetti sociali. *Dedalus*, 1, pp. 27-39.

Han Shan (1993). *The Autobiography and Maxims of Master Han Shan*. En R. Cheung y Ch. Yuan Shakya (Traductores). Extraído de http://www.thezensite.com/ZenTeachings.

Le Van Quyen, M. y Petitmengin, C. (2002). Neuronal dynamics and conscious experience: an example of reciprocal causation before epileptic seizures. *Phenomenology and the Cognitive Science*, 1, pp. 169-180.

Lutz, A., Lachaux, J. P., Martinerie, J. y Varela, F. J. (2002). Guiding the study of brain dynamics by using first-person data: synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proceedings of the National Academy of Science*, 99(3), pp. 1586-1591.

Maturana, H. y Varela, F. J. (1985). *The Tree of Knowledge*, Boston: New Science Library.

Reichle, F. (Escritor/Director). (2005). *Monte Grande. What is Life? Francisco Varela* [Documental]. Suiza: Franz Reichle and TyC Film.

Rodríguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B. y Varela, F. J. (1999). Perception's shadow: long-distance synchronization in the human brain. *Nature*, 397, pp. 340-343.

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J. P. y Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco's Varela exploration of the biophysics of being. *Biological Research*, 36, pp. 27-65.

Scharmer, C. O. (2000). *The three gestures of becoming aware. Conversation with Francisco Varela*. Extraído de http://www.appliedecologics.com.

Thompson, E. y Varela, F. J. (2001). Radical embodiment. *Trends in Cognitive Science*, 5(10), pp. 418-425.

Varela, F. J. (1979). Reflexiones sobre la guerra civil chilena. En A. Palacios y A. Cohen-Varela (Eds.) (2011), *La ciencia del ser: las rutas de Francisco Varela*. Valparaíso: Universidad de

Valparaíso Editorial.

—— (1990). On the conceptual skeleton of cognitive science. En Fink W (Ed.), *Beobachter* (pp. 13-25). Munich: Verlag.

—— (1991). Organism: A meshwork of selfless selves. En Tauber (Ed.), *Organism and the Origin of Self*, pp. 79-107. Dordrecht: Kluwer.

—— (1992). *Un know-how per l'etica.* Roma-Bari: Laterza.

—— (1995). Resonant cell assemblies: a new approach to cognitive functions and neuronal synchrony. *Biological Research*, 28, pp. 81-95.

—— (1996). Neurophenomenology: A Methodological remedy to the hard problem. *Journal of Consciousness Studies*, 3, 330-350.

——- (1997). Patterns of Life: Intertwining identity and cognition. *Brain Cognition*, 34, pp. 72-87.

Varela, F. J., Thompson, E., Rosch, E. (1991). *The Embodied Mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

---

[1] El dominio de la deformación a la cual el sistema puede ser expuesto sin perder su identidad (y así, mantener su organización) es el dominio donde existe como unidad.

[2] Sugerimos que el proceso crucial para la conciencia atraviesa las divisones entre cerebro, cuerpo y mundo, más que reducirse a eventos neurológicos en la cabeza.

[3] Estamos en un sistema cognitivo del cual no podemos ni escapar ni elegir dónde inicia o cómo funciona.

[4] Este nivel global del yo como individuo apareciendo en las interacciones relacionadas con la mente es "un modo de existencia del cual no puedes decir que no existes. Francisco no existe".

[5] La cognición es una acción a partir de lo que falta, llenando deficiencias de un yo cognitivo.

[6] "Dejar ir" es un gesto interesante, porque de hecho es casi como invocar la virtualidad del yo, poniéndolo espontáneamente sobre la mesa. Usualmente, es la vida la que te hace dejar ir algo. Tú sabes lo que quiero decir, en casos de extrema enfermedad o peligro, o de desilución amorosa, te fuerza a ese gesto de dejarlo ir, déjalo ir. Es interesante que los seres humanos tienen la capacidad de movilizar esa capacidad todo el tiempo.

[7] ¿Qué significa tener una buena vida? Una vida de sabiduría es estar constantemente en ese "dejar ir", en dejar la virtualidad o fragilidad del yo manifestarse. Cuando estás con alguien que realmente tiene esa capacidad al máximo, te afecta. Cuando conocemos a ese tipo de gente, es claro, porque todo el proceso no es individual, no es privado, y entras en ese tipo de resonancia.

[8] Nacimiento y muerte.

Día y Noche.

Agua corriendo, piscina estancada.

Brote y flor que desaparece.

¿Puedo encontrar un punto en el cual cambian?

¿De uno hacia el otro?

¿Puede mi nariz girar?

[9] Repentinamente me dieron ganas de orinar. Y de nuevo, viendo mi orina correr, todo se detuvo. Experimento el eterno momento, observé con mi ojo Buda. Ahora sé. Nada es nacido y nada puede morir. Todo es simple.

PROBLEMAS ESPECÍFICOS EN FILOSOFÍA DE LA MENTE

CAPÍTULO IV

La crítica de Hans Jonas al monismo materialista

Alejandro Serani Merlo

Introducción

Como suele ocurrir en filosofía, el problema mente-cerebro no es sino una variante contemporánea de un problema multisecular. Esto, desde luego, si se lo considera desde el punto de vista de sus raíces ontológicas fundamentales y no desde los aspectos científico-experimentales, científico-teóricos, epistemológicos o lógicos que con toda razón varían enormemente de época en época. En efecto, desde la perspectiva ontológica, el problema mente-cerebro parece ser una variante moderna de lo planteado en otros tiempos en términos de: cuerpo y alma, carne y espíritu, materia y forma, *res extensa* y *res cogitans* u otras conceptualizaciones dicotómicas o tricotómicas análogas.

En la extensa literatura contemporánea sobre la filosofía de la mente, no son muchos los autores que se dan la libertad de tratar esta cuestión abiertamente. Esta parvedad de estudios tiene diversas explicaciones. Para decirlo de modo resumido digamos que algunos consideran el problema resuelto, otros lo consideran sin interés, y otros no ven siquiera el problema. Entre los pocos autores que han visto el problema, que lo consideran de interés y que no lo han dado *a priori* por resuelto, se encuentra el filósofo judío-alemán Hans Jonas (1903-1993). El neurocientífico chileno, Francisco Varela, valoraba hace pocos años su obra, en la introducción a su último libro, y lamentaba haberla conocido tardíamente:

> Vine a descubrir a Jonas hace pocos años cuando las ideas de este libro ya estaban asentadas. Con pasión encontré una filosofía de la biología que daba una amplitud a precisamente *todo* el fenómeno de la vida hasta ahora inigualado. Por lo mismo este pensador, todavía poco conocido,

me dio una confianza en mis propias intuiciones. Lo traigo aquí como figura tutelar que sigue siendo fuente de inspiración (Varela, 2010).

Los planteamientos de Jonas acerca de este tema, se encuentran en su más clara y sintética expresión, en un breve texto titulado “El materialismo, el determinismo y la mente”, incluido como un apéndice a un capítulo en su libro *The phenomenon of life, toward a philosophical biology* (Jonas, 1982). Este libro fue traducido por el mismo Jonas al alemán como *Das Prinzip Leben*, por lo que la versión en español, traducida a su vez del alemán, lleva por título *El principio vida, hacia una biología filosófica*. El apéndice es un texto corto y denso, de carácter fundamentalmente crítico: más que mostrar positivamente lo que piensa el autor sobre el tema, intenta mostrar las inconsistencias de la posición criticada.

La biofilosofía de Jonas

Antes de entrar de lleno en la crítica de Jonas al modo dominante de plantear el problema mente-cerebro, sobre todo en el mundo anglosajón, conviene situar previamente sus juicios críticos en el contexto más amplio de su reflexión sobre los seres vivos.

Sobre la base de una amplia panorámica histórica, Jonas piensa que la concepción que los seres humanos se hagan de los seres vivos determinará la ontología subyacente. En definitiva, lo que una filosofía tenga que decir acerca de los seres vivos manifestará sus posiciones metafísicas centrales. Sostiene el autor que la cultura occidental ha basculado históricamente entre una visión panvitalista de la naturaleza –propia del mundo antiguo y de la filosofía naciente–, y una visión panmecanicista, que en sus planteamientos contemporáneos no sería sino el producto de descomposición del dualismo cartesiano. Dualismo que se manifestó durante un interludio breve pero significativo de la historia del pensamiento occidental. De allí la caracterización que hace Jonas del pensamiento imperante en las discusiones científico-filosóficas actuales, como de un mecanicismo postdualista (Jonas, 2000).

Ni panvitalista ni panmecanicista, ni dualista ni monista absoluto, Jonas intenta una vía media que él califica de “monismo integral”, que pretende no desconocer ni la originalidad de lo vivo ni las prerrogativas de la

materia. El metabolismo, la percepción sensible y el conocimiento racional serían para Jonas, en este contexto, realizaciones graduales de lo viviente e irreductibles a la mera causalidad eficiente material, operativa solamente en el espacio físico. La vida no es, para este autor, una suerte de correlato mental de la actividad físico-química, sino definitivamente "otra cosa". No sería adecuada entonces la conceptualización de la vida como un resultado acumulativo de actividades físico-químicas azarosa y felizmente concertadas, sino, por el contrario, la vida sería la razón y la causa de la integración de las actividades físico-químicas. Esta confluencia o integración de dinamismos es la que redunda, según Jonas, en la constitución de un límite ontológico que separa lo inerte de lo vivo. Libertad del viviente por relación a las fuerzas disgregadoras del entorno, constitución de un límite entre el vivo y su medio, instauración de grados progresivos de mediatez entre lo vivo y lo inerte, aparición de una "interioridad" correspondiente a cada uno de esos grados o niveles ascendentes de mediatez, son para Jonas las principales notas distintivas de lo vivo (Serani y Lailhacar, 2013).

La crítica jonasiana al materialismo postdualista

Es desde esta base brevemente esbozada, que Jonas se plantea críticamente frente a darwinistas, conductistas y cibernéticos en el texto que vamos a examinar, y desde donde enrostra a estas posiciones un cartesianismo implícito que no se hace cargo de "el lastre metafísico de su doctrina de las dos substancias" (Jonas, 2000, p.182). ¿A qué se refiere Jonas con ese "lastre metafísico" de la doctrina de las dos substancias?

Según Jonas, el positivismo que hunde sus raíces en el dualismo de Descartes, asumió una sola de las dos ramas de la bifurcación entre substancia pensante (*res cogitans)* y substancia extensa *(res extensa)*. Según esto, el edificio ideológico de la ciencia experimental se habría construido haciendo caso omiso de la *res cogitans,* pretendiendo que esta era una derivación espiritualista o teológica específica de Descartes, la cual no era imprescindible para hacer una ciencia de la naturaleza. No es que se nieguen los atributos espirituales en sí mismos, dice Jonas, como ocurría en los materialismos predualistas de la antigüedad, sino que se trata de

mantenerlos alejados de los datos físicos. Para el darwinismo por ejemplo, los atributos mentales serían una suerte de "lujos" generados por la mecánica de las mutaciones orgánicas. En efecto, solo cuentan en la mecánica de la selección, los aspectos conductuales que inciden en una sobrevivencia diferencial. Solo esos aspectos observables en su "exterioridad" podrían ser tratados por la ciencia (Jonas, 2000, p. 182; Jonas, 1982, p. 127).

Es posible, e incluso probable, que ciertas acciones animales vayan acompañadas por estados de conciencia, pero de conformidad con el axioma de que los procesos exteriores deben ser explicados solamente con arreglo a normas exteriores, es decir, físicas, esos estados carecen de toda relevancia científica. Si la cadena de estimulación sensible, conductos nerviosos aferentes, sinapsis centrales, conductos eferentes y excitación muscular se puede reconstruir sin saltos y lagunas, no hay necesidad ni sitio para interpolar lo mental como un eslabón en algún lugar de la cadena, por mucho que se admita su aparición secundaria en algún lugar a lo largo de la cadena. Superfluo en el cómputo dinámico y sin que ni siquiera sea un dato observable, lo mental queda excluido del universo de la descripción científica. Gracias a esta parsimonia, el behaviorismo consigue disfrutar de las ventajas de la posición cartesiana sin tener que cargar con los inconvenientes del dualismo propio de esta última. Sin embargo, la doctrina de que lo mental es superfluo desde el punto de vista causal es por su parte una posición metafísica que el behaviorismo, mal que le pese, lleva consigo al menos implícitamente (Jonas, 1982; Jonas, 2000).

Esta omisión de la vida mental, piensa nuestro autor, dejó pendiente la tarea de dar un soporte intelectual a toda una parte de la realidad, "como un intento de hacerse cargo de ello lo mental comparece en función del cuerpo de modo enteramente unilateral, es decir sin reciprocidad" (Jonas, 2000, p. 183). Es así como muchos autores modernos afirman sin más que lo mental es producido por lo material, pero que lo mental no revierte ni causal ni significativamente sobre lo material. Sin embargo, Jonas constata que esta unilateral dependencia que lo mental tiene de lo corporal, no tiene cabida en el esquema de la correlación cuantitativa entre causa y efecto, y aparece, por lo tanto, como ideada *ad hoc* para defender los intereses de la ciencia natural, manteniendo así el principio derivado del cartesianismo del cierre o clausura causal del ámbito material. Manteniendo el papel de "causa de lo

mental” que ahora se hace jugar a la materia, pero haciendo a la vez que este rol causal sea tan periférico a la materia como sea necesario, a fin de no comprometer la pureza de su concepto, surge de modo *ad hoc*, dice nuestro autor, la teoría de lo mental como epifenómeno de la materia. Epifenómeno concebido como un producto secundario de procesos cerebrales regidos por leyes estrictamente materiales (Jonas, 2000, p. 183).

Lo más relevante del “epifenomenismo” –continúa Jonas– son las tesis negativas que oculta. La rotunda y afirmativa tesis de que la materia es responsable de la mente, es proferida sin ningún intento de mostrar cómo un tal desempeño podría sustentarse en las propiedades que conocemos de la materia. Esta no es más que una aseveración de lo oculto. Aspira a designar un efecto que, a diferencia de todos los demás efectos en la naturaleza, no consume la energía de su causa: no es una transformación o una prolongación de una tal energía, por lo que, a diferencia de todos los otros efectos conocidos, este no puede ser por su parte una causa. Es impotente, en sentido absoluto, un callejón sin salida que se desvía del camino real de la causalidad, y por cuya entrada el tráfico de las causas y efectos pasa de largo como si no existiese (Jonas, 2000).

A través de una serie de expresiones como “fulguración”, “emergencia”, “epifenómeno”, que más tienen de metáfora que de verdadera explicación, se pretendería afirmar un modo *sui generis* de causación acerca del cual no se tiene ninguna idea positiva. Como dice Jonas, la resbaladiza fórmula del epifenómeno delata este estado de cosas precisamente por el hecho de que intenta ocultarlo. Así, se consigue cerrar el sistema de la causalidad material con la misma eficacia lograda por el dualismo cartesiano. Pero el epifenomenismo dejando aparte el dispendio tan grandioso como inútil en que habría incurrido la naturaleza al permitirse el lujo de la conciencia, trae consigo dificultades, e incluso absurdidades, que forzosamente tienen que aparecer como un precio demasiado caro que pagar por la comodidad científica a cuyo servicio fue ideado (Jonas, 2000).

En efecto, si la vida, la percepción y la conciencia son solo epifenómenos que no revierten causalmente sobre los fenómenos que los originan, entonces se trata de realidades superfluas. Si los seres vivos son máquinas como quieren Descartes y su numeroso séquito de autores contemporáneos, todo su comportamiento se explica en función de su operar físico-químico: ¿qué objeto puede tener entonces para ellos el

conocer, tener sentimientos y pensar? Jonas va aún más allá, en esa hipótesis maquinista, el mismo argumento mecanicista se autodestruye ya que es un argumento enunciado por una máquina, cuyo enunciar argumentos no tiene relación ninguna de correspondencia con lo físico. Si pensar es superfluo, el mismo argumentar es superfluo.

Pero más allá de la reducción al absurdo en que concluye el epifenomenismo, Jonas sostiene que el argumento mecanicista no es ni siquiera coherente consigo mismo:

> Un punto que se suele pasar por alto es que una "materia" que tiene que responder de la mente ya no es la misma materia que la ciencia natural asumió tras la purga dualista. El materialismo heredó su planteamiento del dualismo, sin tomar plena conciencia que el abandono en que incurría comportaba una obligación que nunca podría esperar satisfacer con sus propios recursos: la obligación de asumir teóricamente también aquellos fenómenos que eran la responsabilidad de aquella parte del dualismo que desapareció. Esta tarea se ha vuelto de modo inescapable sobre el materialismo, una vez que se encontró establecido en su parcela, como un monismo auto-suficiente. El peso de esta obligación se ha constituido en la secreta venganza por haber llevado a cabo una de las más grandes usurpaciones en la historia del pensamiento. En realidad, el materialismo no podrá jamás, en virtud de su nacimiento, alcanzar el estado de un legítimo monismo. Esto en razón de que él representa solo una parte aislada de un dualismo que había previamente cercenado una unidad preexistente. El materialismo continúa presuponiendo "lógicamente" al dualismo trascendental, porque solo teniendo de modo inconfesado en la trastienda al "otro mundo" del dualismo puede el materialismo, en su propio campo, pretender desconsiderar la evidencia mental e interpretar la realidad, en términos de pura materia… El mero materialismo, por lo tanto, vivió de su implícita carga dualista y su concesión sobre la vida expiró con la renuncia al complemento espiritualista. Despojada de la coartada dualista, la "materia" solitaria debe dar cuenta ahora de la mente y pierde por lo tanto su naturaleza inambigua de "mera materia" como se la concibió en su momento (Jonas, 2000, p. 184).

Mientras que el monismo materialista de la *res extensa* conserva así la aptitud metodológica para la ciencia natural que había adquirido en el dualismo cartesiano, solo puede evitar el callejón sin salida metafísico de este último al precio de destruir toda posible comprensión de la mente, aun destruyendo la propia idea de mente; y al mismo tiempo, como hemos visto, llena de impurezas el límpido concepto de materia, sobrecargándola de una facultad oculta, esto es, la de generar el "epifenómeno".

En síntesis, la crítica de Jonas a la consideración imperante, que hace de

los fenómenos mentales una suerte de secreción epifenoménica del cerebro, se basa en que no se sostiene racionalmente. En efecto, la consideración postcartesiana de la materia puramente física, no admite el agregado *ad hoc* del epifenomenismo, ni tampoco da cuenta de los fenómenos más significativos de la realidad natural como son la originalidad irreductible de la vida, la percepción y el pensamiento. Por lo primero el epifenomenismo materialista sería incoherente; por lo segundo, irreal.

Analizando a Jonas

Los planteamientos críticos de Jonas, a lo que el designa como monismo materialista, no parecen derivar solamente de la constatación de las inconsistencias lógicas y de las reducciones al absurdo que hemos reseñado, sino que encuentran su raíz última en la conceptualización positiva que Jonas intenta de la realidad natural.

El camino intelectual por el cual Jonas llega a sus planteamientos biofilosóficos y bioéticos de la etapa final de su vida, es de lo más interesante. En efecto, ¿cómo un filósofo formado en la cantera de Husserl y Heidegger, y luego especializado en el análisis histórico de las religiones gnósticas ha llegado a un planteamiento crítico del positivismo cientificista y a una conciencia lúcida de los peligros ínsitos en el dinamismo de la técnica contemporánea? Los quince años pasados como colono judío en Palestina y como soldado de dos guerras no han sido ciertamente ajenos a este devenir intelectual.

Por una parte, el alejamiento afectivo e intelectual de su maestro Heidegger es seguido de una crítica al nihilismo metafísico y moral que, según él, se sigue inevitablemente de este pensamiento. Por otra parte, la experiencia brutal de una guerra donde por primera vez en la historia de la humanidad se desarrollan y aplican armas de destrucción masiva contra la población civil, y donde además la ciencia y la técnica son puestas al servicio de una programa político inhumano, lo alertan acerca de la necesidad de una dirección ética del comportamiento en el plano político colectivo. Más aún, lo ocurrido en Alemania, en los Estados Unidos de Norteamérica y en la Unión Soviética con el desarrollo de la física y de la química y su aplicación a la sociedad y a la guerra, lo alertan acerca de los

peligros potenciales que se avecinan en razón del progreso de las ciencias biológicas y lo hacen interesarse genuinamente en el desarrollo histórico de las ciencias naturales en la modernidad. Finalmente, la tranquilidad que encuentra en la edad madura de su vida para desarrollar con relativa serenidad su tarea filosófica, le permite profundizar en el conocimiento amplio de la historia de la filosofía, desde la antigüedad griega hasta los autores modernos y contemporáneos.

Todo lo anterior confluye en el particular estilo de su filosofar. Fuera de las pugnas ideológicas, Jonas emprende una tarea de comprensión amplia y profunda de la situación cultural y social que lo rodea, y con libertad de espíritu no trepida en inmiscuirse en cuestiones políticas y bioéticas candentes, como la muerte encefálica, las técnicas de reproducción artificial, la clonación de seres humanos o el tema de la ecología. En su filosofía destacan por un lado su manejo del método fenomenológico, y por otro lado el acercamiento progresivo hacia posiciones metafísicas y éticas de carácter realista.

Desde el punto de vista biofilosófico y epistemológico, que nos interesa, Jonas es consciente que la conceptualización que se lleva a cabo desde el plano científico (cientificismo) y posteriormente también desde el plano técnico (tecnicismo), no es ni filosóficamente neutra, ni inocua desde el punto de vista cultural. La conceptualización científica introduce mediaciones racionales que siendo legítimas en su campo, al ser asumidas filosóficamente ocultan y a veces hasta tergiversan nuestro acercamiento al carácter primario de la realidad.

No contento con reconocer y denunciar estas mediaciones, se ha interesado por rastrear su origen ideológico, mostrando como el contexto filosófico en el que estas surgen determina en buena medida las formas que adoptan y las confusiones latentes que vehiculan. Esto aparece con claridad cuando examina por ejemplo el desarrollo histórico de las conceptualizaciones físico-matemáticas o el surgimiento del darwinismo como fenómeno intelectual.

Siendo particularmente sensible intelectualmente al carácter original e irreductible de los seres vivos y de sus manifestaciones plurales y jerárquicas, desarrolla una original biofilosofía, en la que su virtuosismo fenomenológico encuentra una feliz aplicación. Su opción latente, aunque no desarrollada, por posturas metafísicas de carácter realista lo llevan a

percatarse de las encrucijadas ontológicas y de los desafíos éticos que subyacen a ellas.

En el breve análisis que hemos hecho sobre sus posturas críticas de ciertos planteamientos contemporáneos en relación al estatuto de la realidad mental, hemos visto cómo el discernimiento epistemológico de las afirmaciones que derivan de estas posturas, supuestamente avaladas por el desarrollo científico, le permiten detectar con lucidez y fuerza las inconsistencias y o absurdos más o menos habilidosamente ocultos en la narrativa estándar contemporánea.

Queda por determinar en otro trabajo (Serani y Lailhacar, 2013), qué solución encuentran estos dilemas reales en su propia biofilosofía. Esto último teniendo en cuenta sin embargo las prevenciones que él mismo hacía en la introducción a su obra capital *The phenomenon of life*:

> Nuestros análisis… no ofrecen teoría alguna cerrada… si bien es verdad que el autor tuvo a la vista una teoría de ese tipo a modo de objetivo… las distintas investigaciones –así quiero pensarlo– expresan un punto de vista común y constituyen diversas facetas de una filosofía del organismo y de la vida que está por terminar. El autor no se atreve todavía a presentarla sistemáticamente, pero su exposición menos comprometida… puede proporcionar una cierta idea de su configuración, aún en devenir, y recorre ya algunas de las etapas del camino que finalmente quizá pueda conducir a ella (Jonas, 2000, p. 19).

Referencias

Jonas H. (1982). *The Phenomenon of life: toward a biological philosophy*, Chicago: The University of Chicago Press. Reprint: Originally published (1966). New York: Harper y Row.

——(2000). *El principio vida: hacia una biología filosófica*. Madrid: Trotta traducida de *Das Prizip Leben: Ansatzen su einer philosophischen Biologie.* Frankfurt am Mein y Lepizig: Insel Verlag.

Serani Merlo A., Lailhacar Formigo Y. (2013). *La conducta animal y lo transanimal en el hombre según Hans Jonas.* Trabajo presentado en Coloquio Internacional *La fenomenología de la vida de Hans Jonas*, septiembre, Facultad de Filosofía, Pontificia Universidad Católica de Chile.

Varela, F. (2010). *El fenómeno de la vida.* Santiago de Chile: J. C. Sáez Editor.

CAPÍTULO V

Género, imitación e inteligencia: una revisión crítica del enfoque funcionalista de Alan Turing

Rodrigo A. González Fernández

Introducción

Desde que el juego de la imitación vio la luz (Turing, 1950), renació una importante discusión en filosofía de la mente: ¿Es posible en principio la inteligencia de máquina? Tal como analizo aquí, la simplificación de la versión estándar de dicho juego ha tenido un impacto negativo sobre la apreciación de la teoría de la mente que inspira a Turing, lo que a su vez ha eclipsado su proyecto funcionalista del niño-robot. Este trabajo, dividido en cuatro secciones, revisa el Test de Turing, mostrando las dificultades de la versión estándar, las virtudes de la versión de género y el problema al que lleva el proyecto funcionalista del niño-robot.

Justamente, en la primera sección del ensayo se discute el objetivo central del juego de la imitación, en función de la disputa conceptual de la inteligencia de máquina heredada a partir de Descartes, quien descarta de plano la posibilidad de la inteligencia de máquina. En la segunda sección se critica cómo la versión estándar del juego omite del todo el género. En relación con este punto, argumento que, entendido en sentido minimalista como el sexo de los participantes, es crucial para comprender qué tiene en mente Turing con imitar, y también para ponderar correctamente su proyecto: la programación, el aprendizaje y la educación como medios para crear inteligencia de máquina. En la tercera sección se ofrecen razones sobre por qué Turing se basa en la imitación, relacionándola con cómo la inteligencia de mujeres y hombres puede aprenderse por parte de máquinas programadas, de un modo análogo al que lo harían niños-robots.

Finalmente, en la última sección, se critica cómo la imitación de inteligencia de la forma hacerse-pasar-por, fundamento tanto del test como del proyecto del niño-robot, es insuficiente para la acción inteligente en la ontología social; en efecto, como argumento, esta no solo requiere conducta inteligente sino además involucrarse-conscientemente-en.

El Test de Turing: un método empírico para poner punto final a un debate conceptual cartesiano

Es claro que la propuesta de Turing se inserta de lleno en la filosofía de la mente. Hay consenso respecto de que su test pretende poner punto final al problema filosófico de si una máquina puede pensar, o puesto de otra manera, si la inteligencia de máquinas es posible en principio. Esta es justamente una pregunta de inspiración cartesiana que ya había provocado una larga y controvertida discusión, la cual ciertamente no fue resuelta antes y, a mi juicio, sigue provocando polémica por la naturaleza del problema examinado. En efecto, Descartes en diversos pasajes sostiene que es imposible en principio que una máquina tenga estados mentales como los de un ser inteligente, con racionalidad. En relación con el nexo entre lenguaje e inteligencia, Descartes (1637) indica que: “una máquina no podría ordenar las palabras de formas diferentes para responder al significado que se dice en su presencia, como incluso el menos inteligente de los humanos puede hacer” (p. 113). Es decir, una máquina podría usar palabras, pero de modo mecánico y automático, sin ser sensible y flexible al contexto y ambiente.

Justamente, para el filósofo francés, animales y máquinas carecen de inteligencia humana, porque solo pueden reaccionar mecánicamente frente a los estímulos ambientales. A diferencia de ambos, el ser humano es capaz de ordenar las palabras en formas diferentes para significar exactamente lo mismo. Esto se debe a que el lenguaje, puesto de una manera simple y breve, implicaría racionalidad según Descartes, y esta, a su vez, sería signo inequívoco de flexibilidad e inteligencia. En consecuencia, el lenguaje contaría como una prueba empírica confiable de la existencia de racionalidad y, por tanto, de inteligencia.

No obstante, también sostiene que la inteligencia también requiere

flexibilidad en la conducta. Con respecto a la relación entre conducta e inteligencia, argumenta lo siguiente:

> Incluso, aunque tales máquinas podrían hacer algunas cosas tal como nosotros les hacemos, o de mejor forma, inevitablemente fallarían en otras, lo cual revelaría que no están actuando de acuerdo con su entendimiento, sino por la pura disposición de sus órganos (Descartes, 1637, p. 113).

Luego, la ausencia de lenguaje y la conducta mecánica en animales representa evidencia empírica de que ellos y las máquinas no son inteligentes a la manera humana. Por supuesto, la calificación de no inteligentes de animales y máquinas está asociada con la distinción metafísica entre dos substancias propuesta por Descartes. De hecho, los animales son máquinas, en la medida que su corporalidad, su *res extensa*, está constituida por complejos mecanismos que reaccionan de manera puramente causal al ambiente. La *res cogitans*, por el contrario, es la razón flexible e inteligente, capaz de producir conducta apropiada para la resolución de problemas complejos, cuestión que no sucede por una mera relación causa-efecto, sino por la participación del entendimiento. Ciertamente, la relación causa-efecto se da entre cuerpo y ambiente, e indica por qué el primero reacciona de modo mecánico, automático y causal de manera totalmente predecible, en función de la constitución biológica, a diferencia de lo que ocurre con el entendimiento.

A la postura que niega mente e inteligencia a animales y máquinas se le ha llamado conductismo cartesiano (Clark, 1982). De acuerdo con este enfoque, la inteligencia es producto de un tipo específico de racionalidad ajena a lo biológico, que además de no poder ser imitada, involucra que los animales carecen de estados mentales intencionales. Tal interpretación de Descartes sin duda ha provocado polémica, por ser no solo contraintuitiva, sino además poco consistente con lo que puede observarse en la propia conducta animal.

Pese a la inspiración cartesiana de un filósofo posterior como de La Mettrie, es sin duda un notable adversario del conductismo cartesiano. Con un estilo provocador, este materialista declara que:

> Ser una máquina y sentir, y ser capaz de distinguir lo correcto de lo incorrecto, como el azul del amarillo –en una palabra, nacer con inteligencia y un instinto certero de moralidad, y ser solo un animal– son así cosas que no son más contradictorias que ser un simio o un loro, y ser capaz de

darse placer a uno mismo (de La Mettrie, 1747, p. 54).

Hubo muchos otros a quienes también impactó el dualismo cartesiano que divide el mundo entre *res cogitans* y *res extensa*, y más aún, el tratamiento de los animales como mecanismos totalmente carentes de inteligencia. En general, la simpatía por el materialismo, asociado al avance de las ciencias, hizo que la filosofía cartesiana fuera sometida progresivamente a crítica. Por ejemplo, un famoso materialista decimonónico, Charles Babbage, quien es considerado uno de los padres de la inteligencia artificial, planteó en diversas ocasiones que no existe imposibilidad en principio de que una máquina emule la inteligencia matemática. Intentó construir dos máquinas para mostrar que las máquinas inteligentes no son imposibles en principio. La máquina de las diferencias, pensada para reemplazar los computadores humanos, y la máquina analítica, diseñada para realizar tareas no solo matemáticas.

A pesar de estos proyectos, la postura de Babbage es cauta cuando se trata de determinar si la máquina analítica en efecto tiene estados mentales, o solo imita la inteligencia matemática. Tómese en consideración la cautela de estas declaraciones en el siguiente pasaje de Swade (2000) en donde alude a Babbage:

> Babbage habla de "*enseñarle a la maquina a prever*". En otras ocasiones afirma que la máquina "*sabe*". El tenía la suficiente perspicacia para saber si este uso era adecuado, y evidentemente sentía que antropomorfizar mecanismos requería de cierta justificación o excusa: La *analogía* entre estos actos y los procesos mentales me forzó al uso *figurativo* de tales términos. El uso de estos fue ponderado como *económico* y *expresivo*, y prefiero seguir usándolos que sustituirlos por largos circunloquios (En Swade, 2000, pp. 103-4, énfasis míos).

En vista del mentalismo cartesiano y de la cautela de Babbage, la estrategia adoptada por Turing en 1950 no consistió en proponer un nuevo argumento basado en la discusión de términos y conceptos como "piensa" y "máquina". Por el contrario, diseñó un test, un método empírico preciso, que fuera capaz de reemplazar la pregunta: ¿Pueden pensar las máquinas? Pero, ¿por qué Turing quiere abandonar la arena conceptual, en consideración de que su discusión entra de lleno en la filosofía de la mente y, en particular, en el problema de las otras mentes?

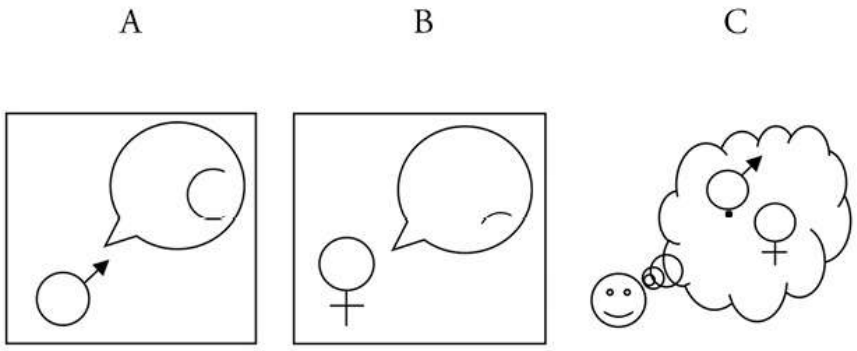
Según él, una buena manera de evitar la pregunta de si las máquinas

piensan consistiría en rechazar el análisis del uso común de conceptos como “piensa” y “máquina”. En efecto, analizar su significado podría llevar a una discusión de definiciones, y así a una encuesta sobre usos de esos términos. El siguiente pasaje del artículo de 1950 es claro respecto de este punto:

> Si la exploración del significado de términos como “máquina” y “piensa” se debe efectuar a partir del análisis de cómo estos se usan regularmente, es difícil evitar la conclusión de que el significado y la respuesta a la pregunta: ¿Pueden pensar las máquinas?, debe encontrarse a través de una investigación estadística similar a una encuesta Gallup (Turing, 1950, p. 40).

Para evitar la discusión conceptual y una eventual encuesta propone un método empírico que reemplace la polémica pregunta. El test se basa en un juego, a saber, el denominado juego de la imitación. Consiste en que hay un hombre en una habitación (A), una mujer en una segunda habitación (B) y un juez cuyo sexo no tiene importancia (C). En rondas sucesivas, el juez formula preguntas sencillas mediante un teletipo, con el objeto de identificar el sexo de los participantes. Las preguntas son simples y orientadas a ese fin, por ejemplo: ¿Tiene usted el pelo largo? ¿Usa maquillaje en algunas ocasiones? Mientras que la mujer responde de manera sincera a las preguntas del juez, el hombre se debe hacer pasar por una mujer. El objetivo del hombre, entonces, es que los jueces se equivoquen y no logren identificarlo correctamente. Así, el juego de la imitación, al menos en su primera etapa, es que un hombre induzca al error en su identificación, y que sea confundido por una mujer. Esta, en cambio, responde de manera sincera. La primera etapa puede esquematizarse así:

Figura 1
Primera etapa del juego de la imitación



Pero, ¿qué tiene que ver el problema del género con la imitación de la inteligencia? Más aún, ¿por qué Turing pasa a la segunda etapa? Mucho se ha elucubrado sobre este cambio, por ejemplo, viendo un proyecto político emancipador de género (Genova, 1994; Lassègue, 1996). A diferencia de estas interpretaciones, estimo que un indicio para entender el cambio a la segunda etapa son las tres preguntas que la motivan:

> ¿Qué sucederá cuando una máquina juegue el rol de A en este juego? ¿Decidirá el juez incorrectamente tan a menudo como cuando se juegue así a como se lo hará cuando en el juego participen un hombre y una mujer? Estas preguntas reemplazan nuestra pregunta original "¿Puede pensar una máquina?"(Turing, 1950, p. 41).

En vista de la forma en que Turing reemplaza la pregunta cartesiana sobre la inteligencia de máquina, al centrarse en la imitación, la siguiente sección examina cómo el juego propuesto por él representa una visión vanguardista con respecto a la inteligencia: es imitable e independiente de la biología, o de cualquier otra realización material.

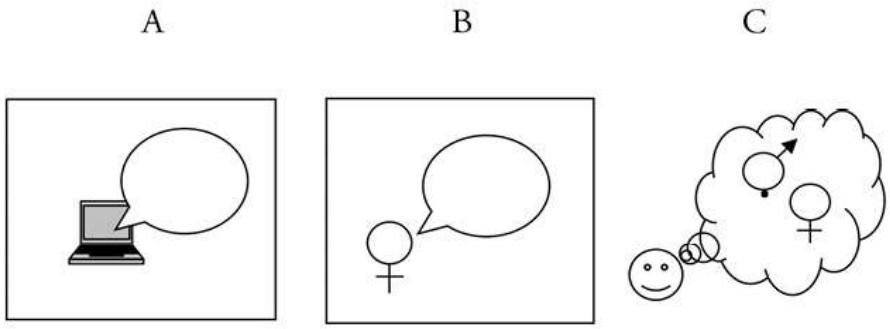
El rol del género en el juego de la imitación: ponderando la justa medida

A diferencia de Descartes, Turing estima que existe evidencia para aseverar que una máquina tiene estados mentales y es inteligente. Por ello es un defensor de la llamada inteligencia de máquina, concepto basal de la inteligencia artificial según la cual el pensamiento mecanizado sí es posible en principio. No obstante, y tal como se analizará con mayor profundidad en la tercera sección, no es necesario que una máquina tenga propiedades biológicas específicas, o que tenga cerebro para que sea inteligente. Por el

contrario, la inteligencia sería una propiedad mental ajena a las propiedades físicas de un sistema cognitivo y así, a lo biológico.

Para comprender el enfoque funcionalista y anti-biológico de Turing, y especialmente su propuesta vanguardista sobre la inteligencia de máquina, es necesario examinar con más detención el juego de la imitación y discutir algunas posturas que señalan que lo central de este es precisamente la cuestión del género. En relación con este punto, cabe destacar que el juego no apunta al problema de la identificación de especies, sino que originalmente se centra en el género, esto es, en el sexo de los participantes, una cuestión que no es ni casual ni trivial. Lo anterior queda de manifiesto en la segunda etapa que Turing describe, donde se puede inferir que un computador digital podría reemplazar lo que hace el hombre en A, tal como este esquema muestra:

Figura 2
Segunda etapa del juego de la imitación



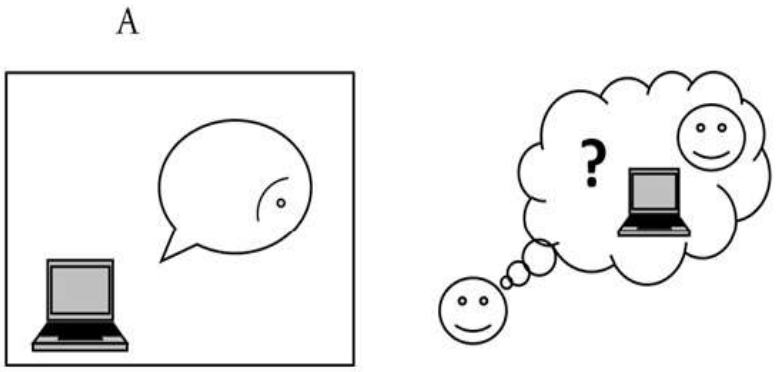
Naur (1986), Hayes y Ford (1995) y Traiger (2003) denominan a esta etapa la “versión de género” del test. Tal como ellos, considero que el género no es una pista falsa [*red herring*] y que no puede ignorarse del todo, tal como algunos aconsejan con la versión estándar (Davidson, 1990). Cabe destacar que los argumentos en pro y en contra de simplificar el test son tan discutidos como la eficacia de este.

Por una parte, los defensores de la versión estándar argumentan que hay cuestiones crípticas en el rol que juega el género en el test, y que Turing describe de modo algo ambiguo que en la etapa 2 la máquina programada

debe jugar “tal como [A]” [*like this*], de la Figura 1 (Shieber, 2004, p. 101). Asimismo, insisten en que en 1952 el propio Turing describe una versión simplificada de su test, dejando de lado la cuestión del género. Tal versión propone el reemplazo de la pregunta sobre si las máquinas piensan así: “Quisiera proponer un tipo particular de test que uno podría aplicar a una máquina […]. La idea del test es que la máquina trate y pretenda ser un hombre” (Turing, 1952 en Copeland, 2003, p. 6). Con base en esta simplificación, se ha tomado la versión estándar como la esencia del test, ignorándose la cuestión del género.

La versión estándar supuestamente cuenta con la ventaja de centrarse solo en la cuestión de la imitación por parte de una máquina, tal como se ilustra a continuación:

Figura 3
Versión estándar del juego de la imitación



Pese a la simplicidad de la versión estándar, parece central en el juego de la imitación cómo un hombre puede imitar la conducta lingüística inteligente de una mujer, y no solo como una máquina puede hacerse pasar por un humano al imitar su conducta lingüística (o como una máquina puede hacerse pasar por cualquier especie). De hecho, según la predicción de Turing, para el año 2000 el computador digital, con más velocidad y capacidad de almacenamiento, debería haber engañado al menos al 30% de los jueces (Turing, 1950, p. 49).

Es importante enfatizar que la simplificación de la versión estándar ha dado lugar a toda clase de interpretaciones incorrectas sobre la filosofía y el

proyecto de largo plazo de Turing. Aunque parece más “simple y neutral” que las etapas 1 y 2, la versión estándar soslaya del todo el género, entendido de manera minimalista: la imitación del sexo de los participantes. Por supuesto, la alusión al género, entendido de manera no minimalista, también ha generado interpretaciones polémicas.

Por ejemplo, Judith Genova (1994) sostiene que podría existir una abierta declaración de Turing a favor de seres inteligentes con nuevas oportunidades de género y sexuales, y esto podría estar inspirado por el pensamiento de Charles Darwin, quien también esperaría la evolución de nuevas formas dentro de un contexto naturalista. Una síntesis de la hipótesis de Genova es la siguiente:

> Genova cree que las nociones sexistas acerca de la mujer siendo menos inteligente, por sí mismas, no explican el diseño particular del juego. Ella sostiene que al complicar el juego de esta manera, Turing cuestiona la existencia de categorías discretas. En otras palabras, al referirse al tópico mujer/hombre, él intenta demostrar que el género mismo es un concepto socialmente impuesto que no es “natural” de la forma en que usualmente pensamos que es. Genova considera el juego de la imitación como parte de la filosofía general de Turing de “transgredir los límites” (Genova, 1994). Bajo el supuesto de que Turing admirara tales transformaciones que no encajan con las categorías discretas, Genova sugiere que Turing podría estar discriminando a la mujer como un ser pensante de categoría inferior por ser incapaz de engañar a un hombre (Saygin, Cicekli y Akman, 2003, p. 59).

Otro ensayo que sostiene otra hipótesis muy discutible, y que se apoya en cuestiones psicológicas y biográficas de Alan Turing[1], es el de Lassègue (1996). Por ejemplo, resulta curioso que la fecha del juicio a que Turing fue sometido por conducta inmoral, a causa de su relación homosexual, haya prácticamente coincidido con la entrevista que dio a *BBC* en 1952, en enero de ese año, y que en ese entonces haya habido preocupación pública por el descubrimiento de dobles agentes homosexuales trabajando para la KGB. ¿Habrá motivado esto que Turing obviara la cuestión del género? ¿Serán solo coincidencias?

Un análisis riguroso del juego de la imitación aconseja no adoptar la versión estándar, que omite el género del todo. Pero tampoco parece aconsejable exagerar tal elemento al extremo de ver un proyecto político. Por el contrario, el género es relevante para caracterizar por qué la imitación es suficiente para recabar evidencia acerca de la inteligencia de

máquina. Cabe recordar que el objetivo de Turing es proponer un método empírico para detener la discusión filosófica de índole cartesiana respecto de la posibilidad de que máquinas piensen y sean inteligentes. La apreciación de este objetivo se refuerza cuando se identifica el *quid* central de dicho juego, a saber, el engaño [*deception*] mediante imitación, criterio decisivo para juzgar la inteligencia al comparar el desempeño de una mujer con hombre y máquina (Saygin *et al.*, 2003, p. 26). Por este motivo, es importante aclarar que hacerse pasar por una mujer (y eventualmente por un hombre) es una propiedad que, al igual que la inteligencia, no requeriría de implementación material específica, como la biológica. Luego, el punto central de Turing es que la imitación de conducta lingüística inteligente en función de un estándar con el cual comparar sería suficiente para determinar, inductivamente[2], si una máquina programada es inteligente, pese a no tener sexo, ni poseer propiedad biológica alguna. En consecuencia, Turing cree que la evidencia observable aportada por el juego de la imitación es suficiente para exorcizar el mentalismo cartesiano y, así, para detener la disputa conceptual al que este da lugar.

El propósito de la siguiente sección es justamente examinar sucintamente qué filosofía habría motivado a Turing a introducir la cuestión del género y su relación con la imitación y la inteligencia, elementos centrales de su concepción anti-biológica de la mente.

La filosofía funcionalista y anti-biológica de Turing: el nexo entre imitación, género e inteligencia

Dos antecedentes serán centrales para sustentar mi posición respecto del juego de la imitación planteado por Turing, quien critica que lo biológico determina crucialmente la inteligencia. Mostraré que los trabajos de Turing de 1936 y de 1950 son los pilares fundamentales del funcionalismo, una teoría de la mente y la inteligencia, y vincularé dicho funcionalismo con un argumento que presenta Turing respecto del aprendizaje y la educación: cómo podría programarse un computador digital a aprender del mismo modo en que un niño se transforma en una mente adulta inteligente, es decir, ocupando mecanismos simples, que guiados por la educación, ayuden a crear aquella (Turing, 1950, p. 62).

En particular, argumentaré que el aprendizaje mediante programación que tiene en mente Turing representa una perspectiva tan vanguardista como polémica respecto de la inteligencia, el funcionalismo, una filosofía anti-biológica.

El funcionalismo de máquina de Turing

Además de ser uno de los padres de la inteligencia artificial, Turing es un precursor de un enfoque filosófico denominado funcionalismo. La tesis central de esta postura filosófica es que la inteligencia es una función no biológica, puesto que lo que interesa respecto de los estados mentales son las relaciones de *input/output* y los estados internos de una máquina. Todo esto puede ser descrito mediante el funcionamiento de una máquina de Turing, el fundamento conceptual para entender qué es una máquina programada o computador digital.

Una máquina programada se caracteriza por tener carácter multipropósito, es decir, por tener la capacidad de resolver problemas de diferente índole. Una máquina de Turing, a su vez, es una idealización matemática de un dispositivo mecánico-abstracto y la base para entender qué es una máquina programada. Junto con operar sobre la base de quíntuplos[3] descriptibles por una tabla (ver Turing, 1936, pp. 231-235), la máquina de Turing puede vincularse con el sentido intuitivo que posee computar (ver Turing, 1936, pp. 250-253). Esta actividad consiste en la ejecución de un algoritmo, cuyos pasos son finitos y calculables mediante lápiz y papel, sin la participación de la conciencia. Para ejecutar un algoritmo la máquina opera de modo mecánico, ya que la tabla dicta el proceder. Tal como sucede con un algoritmo, el término de la ejecución de este ocurre cuando la máquina de Turing logra resolver un problema mediante procedimiento efectivo. Un “computador humano” que emplea lápiz y papel, también puede definirse en términos de una máquina de Turing, o del cálculo Lambda. Este establece operaciones de abstracción y generalización que son realizables por una máquina de Turing, hasta que se alcanza una solución mediante computaciones recursivas (ver Copeland, 1993, p. 230).

Tanto la máquina de Turing como los algoritmos permiten caracterizar

qué es computar. En realidad, esta máquina es una idealización de qué es computar, o de la ejecución de un algoritmo mediante la implementación de un programa. Por este motivo, usualmente se representa a la máquina de Turing mediante una cinta infinita y una cabeza lecto-escritora (o *scanner*) que se mueve hacia la derecha y la izquierda, cambiando de estado interno. Los estados en que se encuentra la máquina de Turing están totalmente condicionados por los símbolos leídos, o estados discretos, y los mencionados estados internos. Por ejemplo, la cabeza se encuentra en un tiempo *t* en un estado interno específico (q_0, q_1, q_2,…, q_n), luego un símbolo de la celda es leído (b_1, b_2, b_3,…, b_n). Dado el estado interno, lo que se lee en la cinta y el micro código de la cabeza, esta mantiene el símbolo, o bien lo borra, imprimiendo otro. Luego, la cabeza para o se mueve a otra celda de la cinta, continuando las computaciones hasta que se detiene. Si esto sucede, se ha llegado al final del procedimiento y, por tanto, a la resolución de un problema.

Tradicionalmente, se representan los símbolos de una máquina de Turing mediante 0s y 1s. Cabe destacar que su conducta está absolutamente condicionada por el programa, descrito mediante la mencionada tabla. Si la máquina está en un estado interno y lee una celda de la cinta, esta dupleta implica un input I_n: <estado interno, símbolo leído>. El programa, entonces, dicta una tripleta de *output* en función del programa: <escribe símbolo, movimiento (o detención) y nuevo estado interno>. Este O_n hace que la máquina escriba un 1 o 0 en una celda, que se mueva izquierda o derecha, ingrese a un nuevo estado, o que se detenga.

Las dupletas de *input* y tripletas de *output* pueden describirse en forma de quíntuplos y enunciados condicionales, todo lo cual ayuda a que se conforme la lista finita de reglas ordenadas de pares <*input*, *output*>. Los quíntuplos tienen la forma <estado actual, leer símbolo, escribir símbolo, moverse (o detenerse) y nuevo estado>. Es pertinente enfatizar que los símbolos de una máquina de Turing pueden representar objetos que no necesariamente sean matemáticos, y esto permite que exista computación de problemas de diverso orden.

Una cuestión crucial respecto de la máquina de Turing es que puede imitar el comportamiento de cualquier otra máquina, sin importar los materiales con que están confeccionadas. Lo único que se requiere es que

ambas funcionen de la misma manera, es decir, que implementen el mismo programa. Así, una máquina de Turing puede imitar la conducta de un cerebro, pero no requiere de la "papilla neural" para tales efectos. Y tal cuestión representa una de las tesis centrales del funcionalismo de máquina de Turing: para tener la inteligencia de una mujer, no es necesario tener sus propiedades físicas o biológicas. Basta con que el funcionamiento de su máquina-cerebro y del computador digital sean similares, tesis de Turing que luego sistematiza y defiende Putnam (1973).

El siguiente extracto de una entrevista radial a Turing en 1951 es iluminadora respecto de la relación entre imitación de inteligencia y funcionalismo:

> Para lograr que nuestro computador imite a una máquina solo es necesario programarlo para que calcule lo que la máquina en cuestión haría bajo ciertas circunstancias […]. Ahora bien, si una máquina en particular puede describirse como un cerebro, tenemos que solamente *programar* nuestro computador digital para imitarlo *y también será un cerebro*. Si se acepta que los cerebros reales, descubiertos en animales, y en especial en el hombre, son una clase de máquina, se seguirá entonces que nuestro computador digital, debidamente programado, *se comportará como un cerebro*. Este argumento presupone una idea que puede ser razonablemente cuestionada […] que esta máquina debiera ser de una naturaleza cuya conducta sea en principio predecible mediante cálculo […]. Nuestro problema es, entonces, cómo programar una máquina para *imitar al cerebro*, o si lo pudiésemos expresar de una manera más breve y menos rigurosa, *para que piense* (Citado en Copeland, 2003, p. 11, énfasis míos).

Aquí se aclara aún más qué tiene Turing en mente en la primera etapa del juego de la imitación, la cual ha originado tanta confusión. Para que un computador o un hombre imiten la conducta inteligente de una mujer, no es necesario que se repliquen ciertas propiedades físicas o biológicas, a saber, aquellas por las cuales se tiene mente y conducta lingüística femenina. Para ponerlo de manera cruda y clara: el punto de Turing es que tal como no es necesario replicar las propiedades físicas de la papilla neural del cerebro para tener la mente y las respuestas inteligentes de este, la primera etapa del juego muestra que tampoco es necesario tener el aparato endocrino y reproductor de una mujer para poseer su mente e inteligencia[4]. Basta que el computador o el hombre imiten la inteligencia de la mujer para tener esta, tal como la mujer podría eventualmente imitar la inteligencia de un hombre; similarmente, no se requeriría poseer las propiedades físicas o biológicas

asociadas al sexo masculino[5]. La pregunta que cabe hacer, entonces, es si tal como la inteligencia, la conducta lingüística orientada por los roles de género puede aprenderse, de un modo similar a lo que Turing plantea con respecto a la creación de la mente adulta. Esto es, mediante la puesta en práctica de mecanismos simples que operen sobre la experiencia y la educación como inputs, que implementen aprendizaje basado en estímulo-respuesta, pero centrado en definiciones y proposiciones expresadas por un lenguaje simbólico, con el objeto de dejar de lado aspectos más bien emocionales (Turing, 1950, p. 63). Si esto fuera factible, la filosofía de Turing y su teoría de la inteligencia serían los pilares de una concepción tan vanguardista como controvertida de la mente, que toca aspectos como el aprendizaje y la educación, todo lo cual haría desestimar que género e inteligencia están puramente determinados por la biología.

Turing, el aprendizaje y la educación: del género al proyecto de crear inteligencia de máquina

Tal como se examinó en la subsección anterior, el juego de la imitación se apoya en una concepción funcionalista de la mente. Puede seguirse de esta propuesta que aprender y ser educado se complementan al crear cierto tipo de inteligencia, la cual no sería un mero producto biológico. Por el contrario, al no importar el material de una máquina que puede imitar la conducta, la inteligencia no está constreñida por lo biológico y lo material. Por esto, su filosofía está lejos de posiciones materialistas contemporáneas, que identifican lo mental con lo neural y que incluso supone que si las mismas propiedades físicas de un cerebro se replican, se tendrá la inteligencia de este. Tómese en consideración, por ejemplo, un experimento mental desarrollado por Block (1990), a propósito de si la inteligencia requiere intencionalidad, la propiedad de algunos estados mentales, como creencias y deseos, de representar y ser acerca de:

> Imagínese un evento con una probabilidad bajísima, pero no cero: dado cierto movimiento aleatorio de moléculas, las partículas de un pantano se juntan y por azar resultan en un duplicado molécula por molécula igual a usted. La criatura del pantano tendrá todas las capacidades (conductuales) que usted tiene, y ellas producirán los mismos procesos psicológicos que ocurren en usted. Luego, podría ser (controvertidamente) inteligente. Pero hay razones para negar que la criatura tiene la

> intencionalidad que usted tiene, y ciertamente para negar que tiene estados intencionales del todo. La criatura del pantano dice, como usted: “Gorvachov influenció a Thatcher cuando viajó a Inglaterra”. Pero, a diferencia de usted, la criatura nunca ha visto a Gorvachov o a Thatcher (o a alguien) en televisión, o ha leído acerca de ellos en los periódicos […]. La criatura del pantano está simplemente *repitiendo palabras*. Si hubiera tenido sus moléculas ligeramente distintas estaría diciendo: “Los sobres Sr. Tattoo Eisenhower están sobre Neptuno” (Block, 1991, p. 263, énfasis mío).

Nótese que aunque Block y Turing coinciden en que hay (controvertidamente) inteligencia cuando hay imitación de esta, las suposiciones materialistas del experimento mental del primero difieren de la filosofía del segundo cuando se trata de qué causa la inteligencia. Mientras que en el contexto del pantano, si hay réplica de propiedades físicas del cerebro, entonces hay inteligencia, según Turing, basta la imitación del comportamiento de la máquina-cerebro para que haya estados mentales e inteligencia. No se requiere, por tanto, del sustrato biológico, y esto hace que para Turing ser inteligente sea algo flexible, alcanzable mediante programación de mecanismos simples que doten a las máquinas de la capacidad de aprender y educarse.

Un pasaje iluminador con respecto a cómo la inteligencia no es un producto físico, y por tanto puede ser programada, aprendida y enseñada es uno en que Turing se refiere al desarrollo futuro de la inteligencia de máquina. Justamente, esta debe tomar en consideración de qué forma aprende un niño, y es educado, para alcanzar la mente adulta. El pasaje en cuestión es el siguiente:

> En el proceso de tratar de imitar a una mente humana estamos condicionados a pensar bastante acerca de los procesos de que fue sujeto para llegar a su estado actual. Podemos notar tres componentes:
>
> 1. el estado inicial de la mente en el nacimiento;
>
> 2. la educación que ha recibido;
>
> 3. otra experiencia, no descriptible como educación, que ha tenido.
>
> En vez de intentar producir un programa para simular la mente humana, por qué no tratamos más bien de simular uno que sea como la de un niño. Si se le sometiera a la educación apropiada, uno obtendría una mente adulta. Tal vez el cerebro del niño es algo así como un libro de notas que uno

compra en una tienda de artículos de oficina. *Pequeños mecanismos y un montón de páginas en blanco.* (Turing, 1990, pp. 61-62, énfasis mío).

Tales mecanismos simples son, precisamente, programables para que trabajen el input de la experiencia, y se logre crear mente adulta inteligente mediante aprendizaje y educación. Nótese que este filósofo no se aboca a dar detalles de cómo debería aprender una máquina a tener la inteligencia de un niño, sino que propone que los mecanismos simples programables son suficientes para activar la cadena aprendizaje-educación-mente adulta. Pero, las preguntas que caben son las siguientes: ¿Puede la inteligencia del niño-robot ser suficiente para la acción inteligente en la ontología social? Más aún, ¿puede el niño-robot de Turing ser inteligente en vez de parecer serlo en dicho contexto?

El problema de la inteligencia de máquina del niño-robot en la ontología social

Estimo que la filosofía de Searle (2010) constituye un sistema. Su concepción de la ontología social explica todos los hechos brutos, de la ciencia, más los institucionales. Aunque es un hecho bruto tener cerebro, este da lugar a todo hecho institucional, declarado mediante lenguaje. Turing, tal como se examina en la sección precedente, está convencido de que la inteligencia es una función, algo programable en máquina de Turing, con independencia de lo biológico. Esto es imposible en la ontología social, como defenderé en el resto de esta sección, pese a la argumentación basada en el niño-robot.

En aras de la claridad y de la economía, es necesario tener presentes los siguientes elementos de la ontología social serleana: cerebro, conciencia, intencionalidad individual y colectiva, realidad física, hechos brutos, hechos institucionales, declaraciones, lenguaje, funciones y marcadores de estatus, reglas constitutivas y poderes deónticos. Brevemente se engarzan estos conceptos de la siguiente forma.

El cerebro, máquina biológica, tiene capacidad de sentir, gracias a la conciencia, y de representar, a causa de la intencionalidad. Tal máquina está inserta en un mundo físico, con el cual se conecta sintiendo. Pero, dada la complejidad de la inteligencia, el cerebro representa objetos y les da

significado, los que aparecen en el ámbito de la ontología social: elecciones, dinero, matrimonios, premios, condenas, etc. Estos objetos descansan sobre propiedades físicas como el papel, las urnas, los anillos de matrimonio, los podios, las cárceles, etc. El hombre, como animal social, se asocia, coopera y participa junto con otros hombres: nace la intencionalidad colectiva; esta origina las atribuciones de funciones de estatus: X cuenta como Y en C. Hago que esta piedra X cuente como mojón fronterizo Y en el contexto de límites entre dos países C. Ciertas reglas se instauran; a saber, se emigra al traspasar la frontera del país propio para habitar territorio extranjero. A partir de esto se originan poderes deónticos: hay una caseta con policías que regulan la entrada y salida por el paso fronterizo. Visten uniformes, sus marcadores de estatus, asociados al reconocimiento de su autoridad. ¿Cómo? Solo puedo salir de la frontera presentando mi pasaporte válido. Pero, aquí viene la parte crucial: estoy consciente de las reglas que regulan la emigración. Para salir de Chile debo poseer un pasaporte. No deambulo de un país a otro. Si emigro, conscientemente presento mi pasaporte y atravieso la frontera a, por ejemplo, Perú. Eso es lo esencial de la acción inteligente emigrar, de involucrarse y hacerse parte de la misma. En vez de eso, los computadores digitales solo siguen reglas formales al modo de un algoritmo que se implementa. Pero la pregunta que cabe es la siguiente: ¿Es la conciencia irrelevante en la ontología social? Por ejemplo, ¿emigraría estando inconsciente, en un estado de catalepsia? Por supuesto que no. Y este punto es central aquí: tampoco lo haría si me transformase en un robot o un zombi. ¿Por qué? El robot sigue las reglas de un algoritmo para implementarlo. Claramente, no hay un sentir que acompañe la ejecución del algoritmo: prueba de esto es que no hay nada que es ser-como-una-máquina-programada. Por otra parte, los zombis por definición también carecen de estados mentales conscientes cualitativos. Por lo tanto, ni zombis ni robots ni hombres con catalepsia sienten y actúan de acuerdo con las reglas constitutivas de “emigrar”. Ese sentir implica estar consciente e involucrarse en las reglas mediante las cuales se participa en acciones de la ontología social.

Y, ¿qué hay entonces del niño-robot y su aprendizaje para convertirse en mente adulta? ¿Sería parte de dicha ontología? Si lo intentase sería una especie de zombi que, al tratar de actuar de acuerdo con reglas constitutivas, produciría acciones mecánicas, imitando la conducta de un

hombre. Aún así, no sería un habitante genuino de la ontología social y aunque estaría en esta no participaría genuinamente en esta. Tal como hay robots que imitan, haciendo como si juegan futbol, los niños-robots harían como-si participasen en la ontología social con acciones inteligentes. Si caben dudas, considérese una orquesta en que se participa colectiva y conscientemente, ¿lo harían los niños-robots de Turing? De la misma forma, ¿podrían tales niños hacer promesas y comprometerse a tocar en su graduación escolar?

La teoría funcionalista de Turing, tan vanguardista como polémica, lleva a un menudo problema en el contexto como la ontología social. No es posible la acción inteligente en un contexto complejo sin un producto biológico como la conciencia, porque de otra forma se actúa como un individuo cataléptico, un zombi o un robot. Es importante finalizar esta sección advirtiendo que no somos zombis ni robots: el hecho de que estemos conscientes de las reglas que constituyen acciones inteligentes en la ontología social hace que seamos los seres más inteligentes del planeta Tierra. También los más peligrosos, por el carácter multipropósito de la inteligencia. Y esta es mucho más multipropósito en la ontología social, en donde se requiere de la conciencia para llevar a cabo la acción inteligente, o flexible, en términos cartesianos. ¿Cuán inteligente? Lo suficiente para navegar exitosamente en el complejo mundo social de funciones de estatus, instituciones y poderes deónticos. Una navegación que requiere más que la pura imitación, o el hacerse-conductualmente-pasar-por, que defiende Turing. Al contrario, requiere de genuina participación o involucrarse-conscientemente-en, condición necesaria de la acción inteligente en la ontología social.

Referencias

Block, N. (1990). “The Computer Model of the Mind”. En E.E. Smith y D.N. Osherson (Eds.) *Thinking: An Invitation to Cognitive Science* (pp. 247-89). Cambridge, Mass.: MIT Press.

Clark, S. (1982). *The Nature of the Beast*. Oxford: OUP.

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

—— (2003). “The Turing Test”. En J. H. Moor (Ed.) *The Turing Test: The Elusive Standard of Artificial Intelligence*, pp. 1-21. Dordrecht: Kluwer.

Davidson, D. (1990). “Turing’s test”. En *Modelling the Mind*, pp. 1-11. Oxford: Clarendon Press.

De La Mettrie, J. O. (1747). *Machine Man*. Reimpreso en S. Shieber (Ed.) *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pp. 45-57. Cambridge, Mass.: MIT Press.

Descartes, R. (1637). *Discurso del Método*. Reimpresión 1984, Madrid: Alianza Editorial.

Genova, J. (1994). “Turing’s sexual guessing game”. *Social Epistemology*, 4, Vol. 8, pp. 313-26.

González, R. (2011). “Descartes, las intuiciones modales y la inteligencia artificial clásica”. *Revista Alpha*, 32, pp. 181-198.

Hayes, P. y Ford, K. (1995). “Turing Test as considered harmful”. *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI 95*, Vol. 1, p. 972.

Hodges, A. (1983). *Alan Turing: the Enigma*. London: Burnett.

Lassègue, J. (1996). “What kind of Turing Test did Turing have in mind?”*Tekhnema*, 3, pp. 37-58.

Moor, J. (1973). “Turing Test”. En S. C. Shapiro (Ed.) *Encyclopedia of Artificial Intelligence* Vol. 2, pp. 1126-30. New York: John Wiley.

Moor, J. (2004). “An Analysis of the Turing Test”. En S. Shieber (Ed.) *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* (pp. 297-306). Cambridge, Mass.: MIT Press.

Naur, P. (1986). “Thinking and Turing’s Test”. *BIT*, 3, pp. 175-187.

Putnam, H. (1973). “The nature of mental states”. Reimpreso en D. Chalmers (Ed.) *Philosophy of Mind: Classical and Contemporary Readings*, pp. 73-79. Oxford: OUP.

Saygin, A., Cicekli, I. y Akman, A. (2003). “Turing Test: 50 years later”. En J. H. Moor (Ed.) *The Turing Test: The Elusive Standard of Artificial Intelligence*, pp. 26-78. London: Kluwer Academic Publishers.

Shieber, S. (2004). “The Ephemera”. En *The Turing Test*: *Verbal Behavior as the Hallmark of Intelligence*, pp. 97-104. Cambridge, Mass.: MIT Press.

Swade, D. (2000). *The Difference Engine: Charles Babbage and the Quest to build the First Computer*. London: Penguin.

Traiger, S. (2003). “Making the right identification in the test”. En J. H. Moor *The Turing Test: The Elusive Standard of Artificial Intelligence*, pp. 99-110. Dordrecht: Kluwer.

Turing, A. (1936). “On computable numbers, with an application to the *Entscheidungsproblem*”. En *Proceedings of the London Mathematical Society*, series 2, Vol. 42, pp. 231-265 (con correcciones en Vol. 43, pp. 544-546).

—— (1950). “Computing intelligence and machinery”. Reimpreso en M. Boden (Ed.) *The Philosophy of Artificial Intelligence*, pp. 40-66. Oxford: OUP.

[1] Supuestamente, la vida de Turing sería relevante con respecto a las teorías que propone. Por ejemplo, se ha interpretado que su condición homosexual tuvo que ver con sus desarrollos teóricos. Para una síntesis de su vida, véase Hodges (1983).

[2] El elemento inductivo del Test de Turing puede apreciarse cuando se sopesa la predicción que él hace: "Creo que en cerca de cincuenta años será posible programar a un computador, con una capacidad de almacenaje de 10^9, para hacerlo participar en el juego de la imitación tan bien que un juez promedio no llegará a tener más de un 70% de probabilidades de hacer la identificación correcta luego de preguntar por cinco minutos" (Turing, 1950, p. 49). Moor (1976 y 1987) y Copeland (2003) han interpretado el Test de Turing en un sentido inductivo, debido a esta predicción. Según ellos, lo que Turing tiene en mente no es definir la inteligencia, ni tampoco proponer condiciones necesarias o suficientes para esta. Por el contrario, el hecho de que participen diferentes jueces en varias rondas, y que en cincuenta años estos no tengan más de 70% de probabilidades de descubrir a la máquina programada, reafirma que es un método empírico para evaluar la inteligencia. En vista de la evidencia inductiva, tanto Moor como Copeland privilegian solo la versión estándar del juego de la imitación. Las razones de esto están expuestas en Copeland 2003, pp. 6-8.

[3] Más abajo me refiero a qué son los quíntuplos, cuando se describe de qué manera opera una máquina de Turing en función de una tabla, o un programa.

[4] Por extraña que parezca esta cuestión, ciertamente se da en el ámbito empírico con el fenómeno de la transexualidad, esto es, de mentes que no corresponden con el cuerpo. Como puede apreciarse, curiosamente hay cierta cercanía entre el dualismo de Descartes y funcionalismo de Turing (González, 2011), pues cuerpo y mente ni son idénticos ni se condicionan mutuamente desde un punto de vista causal. Al contrario, son separables. Debo la sugerencia sobre la transexualidad a Alfredo Minervini.

[5] En relación con esto, difiero de Genova (1994). El funcionalismo anti-biológico de Turing hace pensar que no hay prejuicio contra la mujer, como incapaz de imitar la inteligencia de un hombre. En efecto, dado que imitar la conducta lingüística no requiere de ninguna propiedad biológica, ¿por qué la inteligencia masculina no podría ser imitada por una mujer también? Esto, además, podría ser una pista de por qué Turing abandonó la versión de género en 1952. En efecto, si se comprende el funcionalismo anti-biológico en pos de la inteligencia de máquina, resulta algo engorroso introducir la cuestión del sexo de los participantes. Mi tesis minimalista acerca del género es compatible con tal posición: el sexo y la conducta lingüística inteligente parecen un buen punto de partida para ejemplificar por qué lo biológico es irrelevante. Copeland (2003, pp. 6-7) también defiende una hipótesis sobre el abandono de Turing del tópico del género.

CAPÍTULO VI

Las intenciones como compromisos: un análisis de las posturas de la doble perspectiva sobre la agencia humana

María Ayelén Sánchez

La teoría estándar de la acción: una concepción causalista de la intención

Entre la infinidad de eventos que ocurren en el mundo, algunos involucran movimientos corporales (como cantar una canción, abrir una ventana o estornudar). Algunos de los movimientos corporales son considerados cosas que le ocurren a la persona (temblar o ruborizarse, por ejemplo) pero otros son entendidos como cosas que la persona hace (patear una pelota o servirse un café). Solo a estos últimos los denominamos acciones. El intento por comprender y explicar esta última subclase de eventos motivó el surgimiento de una rama de la filosofía que se ocupa específicamente de reflexionar sobre las acciones. Así nació la filosofía de la acción en la segunda mitad del siglo XX con el objeto de teorizar sobre las acciones intencionales humanas, es decir, aquellos comportamientos motivados por creencias, deseos e intenciones y explicables en estos términos. Esta rama de la filosofía se interesa por las intenciones en tanto que la relación entre la intención y la acción le permite delimitar el campo de las acciones intencionales (Dancy y Sandis, 2015, p. 10).

En la filosofía de la acción se pueden identificar un conjunto de tesis que conforman lo que se denomina "teoría estándar de la acción" (*standard theory of action*), cuyas bases remiten a la obra de Davidson (1963/1980). Sus seguidores han continuado con el espíritu general de la propuesta davidsoniana, ofreciendo
refinamientos, cambios y desarrollos más amplios de sus conceptos fundamentales (Bratman, 2017; Enc, 2003; Mele, 2003; Schlosser, 2011).

La teoría estándar de la acción también suele ser denominada teoría causal de la acción (Aguilar y Buckareff, 2010, p. 9). Un aspecto común a toda teoría causal es que lo que vaya a ser considerado como ϕ, (el tipo de elemento/evento que la teoría intenta definir), no depende de su estructura intrínseca o de sus propiedades, sino más bien de su etiología (una propiedad extrínseca y relacional) (Enc, 2003). El paradigma causal de la teoría estandar de la acción se basa en la relación causal entre eventos (*events-based accounts*). Desde esta perspectiva, el criterio de demarcación entre la acción humana y el resto de los eventos del mundo es que las acciones tienen intenciones, o sus realizaciones neuronales, como sus antecedentes causales (Mele, 2003). Además de este requisito ontológico, la teoría estándar de la acción defiende un requisito epistemológico, a saber, que la ocurrencia de una acción A puede ser explicada apelando a las intenciones por las cuales el agente realiza A, solo si tales intenciones juegan un rol causal en la etiología de la realización de A. En pocas palabras, el éxito epistemológico de la explicación intencional de la acción descansa en una relación ontológica de tipo causal entre la acción y la intención (o sus realizaciones neuronales).

Las distintas teorías que conforman la teoría estándar de la acción coinciden, entonces, en considerar a las intenciones como estados mentales que mantienen una relación de tipo causal con la acción[1]. Desde hace ya varios años, este enfoque causalista/psicologista se ha complejizado, dando lugar a un concepto de intención que no solo opera en el comienzo a la acción, sino que la acompaña a lo largo de su realización. En esta dirección ha sido muy influyente la teoría de Bratman (1987) según la cual las intenciones tienen la estructura y el contenido de planes a futuro. Este enfoque ha influenciado muchas teorías causalistas contemporáneas que reconocen la doble función de las intenciones: causar la acción y guiar y coordinar el comportamiento del agente a través del tiempo (Mele, 2003, 2009; Setiya, 2007). La influencia de Bratman también puede apreciarse en la caracterización sobre la intención que presenta Mele, uno de los defensores de la teoría estándar más populares en la actualidad. En la teoría de Mele, las intenciones tienen una dimensión actitudinal (son actitudes ejecutivas) y una dimensión representacional, siendo su contenido un plan. La caracterización de las intenciones como actitudes ejecutivas hacia planes es lo que explica su rol causal en la producción, mantenimiento y

coordinación de la acción (Mele, 2009).

Desde la teoría estándar de la acción se han establecido tres frentes de debates con otras posturas filosóficas sobre la agencia. En ellos se han enfrentado a la teoría de la causalidad del agente (*agent-causal approach*), a las posturas volicionistas (*volitionist approach*) y a las teorías de la doble perspectiva (*dual standpoint theories*). Los motivo de rechazo a las dos primeras surgen de considerarlas anti-naturalistas, ya que proponen el reconocimiento de nociones irreductibles como es la de agente o la de voluntad. Las posturas de la doble perspectiva, en cambio, han sido rechazadas por la teoría estándar por evadir el verdadero problema metafísico de la agencia, a saber, cómo los agentes pueden hacer un ejercicio de control sobre sus acciones en un mundo cuyos sucesos son explicados en términos de relaciones causales entre eventos (Schlosser, 2011).

Los argumentos esgrimidos en contra de la causalidad del agente y la teoría volicionista han sido exitosos y en general se los encuentra fundamentados en un análisis detallado de las posiciones que intentan rebatir (Mele, 2010; Schlosser, 2011). Sin embargo, lo que queremos mostrar aquí es que no ocurre lo mismo en el caso del rechazo a los enfoques de la doble perspectiva. En primer lugar, las alusiones a este tipo de perspectivas suelen ser muy escuetas en la literatura. Generalmente, se encuentran caracterizadas como afirmando que el ejercicio de la agencia no puede ser explicado desde un punto de vista metafísico sino que debe ser abordado desde un punto de vista normativo (Schlosser, 2015). Con respecto a las intenciones específicamente, se alude a las teorías de la doble perspectiva como aquellas que sostienen que las intenciones no son estados psicológicos sino normativos (Wilson y Shpall, 2016). En lo que respecta a la relación entre la intención y la acción, se les adjudica la tesis de que el vínculo no es causal sino normativo (Setiya, 2018; Schlosser, 2011).

Un aspecto común que atraviesa las teorías de la causalidad de eventos es la fundamentación de sus fortalezas en el hecho de que su concepción de la acción está ubicada dentro de un marco naturalismo[2]. Hay un acuerdo generalizado entre los autores que defienden alguna versión de la teoría estándar de la acción acerca de que sus supuestos ontológicos y epistemológicos permiten dar cuenta de la acción humana considerándola parte del mundo natural en el que ocurren los demás eventos estudiados por

las ciencias fácticas (Enc, 2003; Schlosser, 2011). En este marco naturalista, hay que aclarar que el compromiso con un modelo causalista de la acción no debe identificarse con un compromiso con el determinismo (Mele, 2009). Algunos autores del campo de la psicología han sostenido que el determinismo es la condición necesaria para que la psicología se constituya como una disciplina científica. En esta línea, autores cono Baer (2008), Howard (2008), Nichols (2008) y Bargh y Ferguson (2000) sostienen que sin el supuesto de que las acciones humanas están causalmente determinadas por eventos que las preceden, entonces la investigación en psicología no puede indagar en las causas del comportamiento humano como cualquier otra ciencia lo hace en relación a sus objetos de estudio. Los defensores de la teoría estándar aclaran que sus enfoques no requieren de supuestos tan fuertes como el determinismo. Si las acciones son causadas de manera probabilística o determinista es una cuestión empírica (Mele, 2009). Que las acciones humanas tengan una causa es todo lo que se requiere para que la investigación psicológica sea una empresa viable. Por eso, desde la perspectiva causalista, las condiciones que tienen que darse para que un comportamiento sea una acción es que esta sea producida por una intención o su correlato físico, sin ser esta una relación de determinación necesaria.

En base a este breve panorama podemos resumir la situación en los siguientes términos. La teoría estándar de la acción se presenta a sí misma como la única vía para comprender y explicar las acciones intencionales humanas a partir de la postulación de un vínculo causal entre la intención y la acción. Dada esta perspectiva, se afirma como el único marco teórico en el cual la investigación psicológica es posible, rechazando las demás alternativas por anti-naturalistas o por evadir el problema central de la agencia. Nuestra hipótesis aquí es que esta caracterización de las teorías de la doble perspectiva no le hace justicia a sus propuestas reales. En la literatura de la filosofía de la acción no se encuentra un desarrollo extendido acerca de qué proponen estos enfoques. De esta manera afirmaciones como que las intenciones, su vínculo con la acción y la agencia en general tienen un estatus normativo pueden aparecer como enigmáticas y un tanto oscuras. Además, se suele presentar a las teorías de la doble perspectiva como abordajes que niegan el vínculo causal entre la intención y la acción o que niegan el estatus de la intención como estado

psicológico. Dado este estado de la cuestión, nuestro propósito aquí será doble. Por un lado, intentaremos aclarar qué implica sostener un enfoque normativista sobre la agencia y las intenciones. Por otro lado, mostraremos que diversos autores posicionados en este enfoque han ofrecido un modelo acerca de cómo se conecta causalmente la acción con la intención y que por lo tanto no eluden el problema metafísico de la agencia relativo al ejercicio del control del agente.

Enfoques normativistas de la acción y tipos de normas

El hecho de que la teoría estándar de la acción se oponga al paradigma normativista de las teorías de la doble perspectiva no debe oscurecer el hecho de que los autores defensores de la primera también reconocen la dimensión normativa de las intenciones. Bratman (2007, 2009, 2017), por ejemplo, es un autor que en el marco de la teoría estándar ha desarrollado extensamente los principios normativos a los que están sujetas las intenciones. Incluso en la obra de Davidson (1981) encontramos la idea de que las intenciones son ítems mentales que están sujetos a principios normativos propios de la racionalidad humana. La diferencia entre las perspectivas causales como la teoría estándar y las teorías de la doble perspectiva está en la diferencia entre concebir a las intenciones como estados sujetos a principios normativos y concebirlas como estados normativos en sí mismas (Bilgrami, 2006, p. 232). Pero quizá una diferencia más específica que debe ser señalada es que cuando las teorías causales piensan la normatividad se centran en normas de tipo instrumentales (las que establecen los medios necesarios para alcanzar determinados fines) mientras que las propuestas de la doble perspectiva suelen apelar a normas constitutivas. Las normas constitutivas son aquellas que en cierto sentido crean la actividad que regulan, siendo las normas de juego el ejemplo paradigmático de ellas (Searle 1963).

Otra clasificación relevante entre las normas es la que distingue las normas del ser (*norms of being*) de las normas para la acción (*norms for action*) (Glüer y Wikforss, 2018). Las normas del ser están asociadas con la evaluación, determinando que cierto estado de cosas debe ser considerado valioso en cierto sentido. Las normas para la acción, por su parte, indican lo

que debe hacerse para actuar en conformidad con ellas. Nótese que esta clasificación de las normas ofrecida aquí es parcial (no aspira a ser exhaustiva) y es independiente de la clasificación que podamos hacer de las normas con respecto de su proveniencia, (normas morales, de etiqueta, de prudencia, de juego, etc.). El objetivo de esta clasificación es simplemente señalar que tras la afirmación de que la agencia debe entenderse en un sentido normativo, sucesivas aclaraciones deben ser hechas para comprender qué tipo de normatividad está en juego.

Entre los autores que generalmente son citados como representantes de la doble perspectiva suelen encontrarse Akeel Bilgrami y Christine Korsgaard. Ambas propuestas normativistas comparten como fundamento de sus propuestas normas de tipo evaluativo. A grandes rasgos, su concepción de las intenciones y de la agencia en general se sostiene a partir de considerar la condición de agente racional humano como una condición intrínsecamente valiosa. Por eso sus mayores esfuerzos argumentativos están concentrados en mostrar el valor intrínseco y último de la agencia racional para hacer de ella la fuente de la normatividad. A continuación, haremos un recorrido por las ideas principales de Bilgrami y de Korsgaard para intentar aclarar el sentido de la idea de la agencia como noción normativa, el de la intención como estado normativo y mostrar cómo estas teorías proponen un modelo causal que explica la relación entre intención y acción.

Akeel Bilgrami y la prioridad de las actitudes reactivas evaluativas

Bilgrami es identificado usualmente como uno de los defensores contemporáneos de los enfoques del punto de vista dual. Explícitamente, el autor formula las tesis de que la agencia es un fenómeno normativo y que las intenciones son estados normativos. En esta sección intentaremos dilucidar qué sentido tienen estas afirmaciones en la propuesta de Bilgrami. En primer lugar hay que aclarar que el tema que principalmente le interesa al autor es el auto-conocimiento pero considera que este tópico no se puede abordar fuera de la integración con las nociones de agencia e intencionalidad. La propuesta de Bilgrami se ubica dentro de lo que él mismo entiende como el “giro de Strawson” (*Strawson’s turn*), inspirado en

las ideas desarrolladas por Strawson (1974) en *Freedom and resentment*. No juzgaremos aquí en qué medida la interpretación de Bilgrami sobre las ideas de Strawson es adecuada ni en qué medida se distancia de ella. Simplemente nos limitaremos a caracterizar su perspectiva normativista sobre la agencia y las intenciones.

¿Cuál es esta idea de Strawson que, según Bilgrami, es tan radical como para establecer un giro en la concepción de la agencia? Básicamente, la idea es que las actitudes evaluativas recíprocas tienen prioridad por sobre las consideraciones metafísicas a la hora de caracterizar fenómenos como el de la libertad y la responsabilidad. Esta consideración es extendida por Bilgrami a la agencia y la intencionalidad, conformando el núcleo de su propuesta normativista. La idea central aquí es que las actitudes evaluativas que mantenemos de manera recíproca constituyen nuestra agencia. La noción de agencia es normativa en tanto que los agentes son aquellos sujetos cuyos comportamientos son objeto de actitudes evaluativas como las de culpar, elogiar, premiar y castigar. La normatividad propia de estas actitudes es constitutiva en tal medida que dejaríamos de ser lo que somos como agentes si abandonásemos tales actitudes (Bilgrami, 2006, p. 58). Abandonar tales actitudes sería cometer un "suicidio agencial" (*agential suicide*).

Este tipo de posturas tienen la dificultad de dar cuenta de por qué ser un agente posee un valor intrínseco. Si pudiese explicarse su valor en relación a una característica ulterior a la agencia misma, entonces se retrotraería el problema a explicar por qué ese rasgo es valioso en sí mismo. Además, en este último caso, se daría una razón instrumental en favor de ser un agente. Pero las explicaciones que ofrecen este tipo de posturas no se basan en una concepción instrumental, sino que consideran que dado que la agencia es una noción evaluativa, esta consiste en nuestra reacciones evaluativas recíprocas y hacia nosotros mismos. Por eso "(…) *it is within our own values and evaluations that we can find a reason not to commit normative and rational suicide*"[3] (Bilgrami, 2006, p. 62). Buscar la justificación de por qué ser un agente es algo valioso por fuera de la agencia misma es lo mismo que preguntar qué valor tiene ser una criatura que posee y ejercita valores. Si respondiésemos esta pregunta citando un valor específico, estaríamos justificando la actividad general de evaluar con un valor más específico. No hay ningún valor que justifique la actividad de evaluar misma, justamente

lo que nos constituye como agentes es ese ejercicio de evaluación mutua del que formamos parte. Para desafiar la agencia desde un cuestionamiento de valor, ya estamos teniendo una actitud evaluativa y por lo tanto ejercitándonos en la agencia misma. Ser un agente es ser una criatura que tiene actitudes evaluativas, en este sentido estas actitudes son constitutivas de la agencia misma y no regulaciones independientes del hecho de ser un agente.

Desde esta concepción normativista de la agencia, Bilgrami deriva su caracterización de las intenciones (y los estados intencionales en general), como estados normativos. Lo propio del agente es poseer estados intencionales. Estos estados son concebidos en términos normativos dado que marcan una diferencia en el mundo a través de nuestras acciones en un sentido que no es causal ni disposicional como se lo entiende en la literatura causalista (Bilgrami, 2006, p. 87). La concepción de las intenciones que está a la base de este enfoque es que los estados intencionales son compromisos (*commitments*) o deberes internos (*internal oughts*).

Todos los estados intencionales pueden ser entendidos como disposiciones. Pero Bilgrami argumenta que este sentido no agota la definición de estado intencional. Para mostrar esto apela a un argumento mooreano, señalando que siempre que existe un estado intencional ϕ, entendido como disposición, permanece la pregunta abierta de si el agente debe realizar la acción correspondiente (p. 213). Tener un estado intencional ϕ compromete al agente a tener otros estados intencionales que se encuentran en relación con él. Este requerimiento normativo es independiente de si el agente efectivamente tiene la disposición a adoptar tales estados intencionales y esta noción de compromiso no se explica desde la mera concepción de los estados intencionales como disposiciones (p. 213). Finalmente, las normas de tipo evaluativo tienen un rol central en esta concepción normativista de las intenciones. Si un agente realiza una acción A motivado por su intención ϕ, queda pendiente la cuestión evaluativa de si la acción satisface o no la intención. Si la intención fuese simplemente causa de la acción, no habría espacio para que la acción no satisfaga la intención.

Bilgrami especifica que dentro de las condiciones de tener un compromiso está la de que uno debe estar preparado a tener ciertas actitudes reactivas, como ser autocrítico o a aceptar críticas de terceros, si

uno no cumple con su compromiso o si uno carece de la disposición a hacerlo. Esto se complementa con la exigencia de hacer lo que esté al alcance de uno para cumplir con el compromiso (p. 226). Esta declaración de Bilgrami se torna especialmente relevante ya que exhibe los elementos centrales de su postura. En primer lugar, un estado intencional es un estado normativo porque establece aquellas acciones que debo hacer en orden a satisfacer la intención. En segundo lugar, tener una determinada intención es estar bajo el influjo de una normatividad que implica normas de tipo evaluativo. En esta concepción de los compromisos, lo que ellos generan son el derecho a adoptar ciertas actitudes reactivas evaluativas, derecho adquirido por terceros como por el propio agente. Finalmente, este enfoque normativista se diferencia del disposicionalista ya que no establece que cuando uno tiene una intención adquiere la disposición a actuar de determinado modo y a tener otros estados intencionales, sino que este principio señala lo que uno debería hacer en tanto agente racional.

Finalmente, resta por aclarar en qué sentido Bilgrami se declara anti-causalista y anti-disposicionalista. El autor aclara que todo lo que requiere una posición normativista es que haya un aspecto de la mente humana y de la agencia que no se explique apelando a disposiciones (Bilgrami, 2006, p. 207). La tesis de Bilgrami es que los estados intencionales no son enteramente disposiciones y su relación con la acción no se agota en ser una relación causal. Ahora bien, ¿se aplica aquí el cargo de que las posturas normativistas dejan a un lado las cuestiones metafísicas y no pueden dar cuenta de la relación entre las intenciones y las acciones? La respuesta es negativa en lo que respecta a los desarrollos de Bilgrami. El autor declara estar interesado en desarrollar una metafísica de la agencia basada en normas (*norm-based metaphysics of agency*). Esta postura no elude el problema metafísico desentendiéndose del problema causal de la intención, sino que afirma que no puede entenderse la cuestión causal dejando a un lado el aspecto normativo de la intención (p. 53). La clave de esta metafísica basada en normas es la idea de que las intenciones tienen su eficacia causal en tanto que compromisos. Los agentes racionales, como los agentes humanos, solemos actuar en conformidad con nuestros compromisos, (y, en particular, con nuestras intenciones). Explica Bilgrami: “*They will have to take the form of saying not that someone’s commitment that –or to– x caused him to do y, but that this commitment to and caused*

him to try and live up to the commitment to do x"[4] (2006, p. 243). Por eso, la eficacia causal de las intenciones descansa en las disposiciones de segundo orden que adoptamos ante los compromisos al tratar de cumplirlos (p. 243). Estas disposiciones de segundo orden son objeto de regulaciones de las normas evaluativas que aplicamos de manera recíproca a otros y a nosotros mismos.

Christine Korsgaard y la identidad práctica como fuente de normatividad

Christine Korsgaard es otra de las fervientes defensoras de la tesis de que la agencia humana es esencialmente normativa. Al igual que Bilgrami, hace de las normas valorativas la piedra angular de su argumentación. El concepto fundamental en el enfoque de Korsgaard es el de la identidad práctica y su carácter normativo. La identidad práctica es el modo en que los agentes humanos se conciben a sí mismos. Obviamente, esta encierra una multiplicidad, ya que una misma persona puede concebirse como ciudadano, como padre o madre, como practicante de alguna religión, como persona honesta, etc. Las dos ideas centrales que defiende la autora pueden resumirse de la siguiente manera: la identidad práctica es normativa y es condición de posibilidad de nuestra agencia racional.

El carácter normativo de nuestra identidad se muestra en el hecho de que ella es fuente de principios y valores (Korsgaard, 1996). Tales principios y valores regulan nuestras actitudes de segundo orden que tomamos hacia nuestros estados intencionales de primer orden. Por ejemplo, si una persona tiene el deseo de alistarse en el ejército para pelear por su país en la guerra, deberá considerar si actuar o no en conformidad con ese deseo. A la hora de tomar esa decisión, estará condicionado por su autoconcepción. Si se considera una persona patriota, eso contará como una razón para realizar su deseo. Si en cambio considera que su identidad práctica como padre de familia requiere no alejarse tanto tiempo de sus hijos, no actuará en consonancia con ese deseo. Así la identidad práctica es fuente de autoridad, y determina las valoraciones y obligaciones para cada agente.

La identidad práctica es, en tanto que normativa, condición de posibilidad para el ejercicio de nuestra agencia humana. Los agentes

humanos no somos "*wantons*"[5], criaturas que actúan de manera directa impulsadas por sus deseos, inclinaciones e intenciones de primer orden. Nuestra capacidad reflexiva nos permite tomar distancia de nuestros estados intencionales y decidir actuar o no en función de ellos (p. 96). La idea de Korsgaard es que sin identidad práctica no tendríamos criterio para evaluar nuestras disposiciones y deseos y careceríamos de un marco para decidir cómo actuar. Por eso la agencia humana es normativa y su ejercicio es la expresión de nuestros principios y valores dados por nuestra identidad (p. 116).

De esta manera, aunque sea una cuestión contingente que identidad práctica adopta cada agente e incluso puedan cambiar en el transcurso de una vida, lo que es necesario es que todo agente humano posea una identidad práctica. La argumentación presentada por Korsgaard culmina con un argumento de tipo trascendental. El problema que enfrenta la autora es el de mostrar que, puesto que toda identidad práctica normativa es fuente de valoraciones y por lo tanto de nuestros cursos de acción elegidos, ser un agente humano tiene un valor intrínseco no instrumental. Para Korsgaard, el valor de ser un agente humano solo puede apreciarse desde una perspectiva de primera persona (p. 124). Ser un agente humano es alguien que ejerce valoraciones sobre posibles cursos de acción. Al adoptar una actitud valorativa estamos al mismo tiempo reivindicando nuestra condición como agentes capaces de valorar, y por lo tanto, cuando valoramos algo valoramos a su vez nuestro carácter como agentes humano. En esta dirección, la autora sostiene: "*Most of the time, our reasons for action spring from our more contingent and local identities. But part of the normative force of those reasons springs from the value we place on ourselves as human beings who need such identities*"[6] (p. 121) y más adelante agrega: "*Since you cannot act without reasons and your humanity is the source of your reasons, you must value your own humanity if you are to act at all*"[7] (p. 124). Por esta vía la Korsgaard intenta fundamentar que la agencia humana posee un valor intrínseco y último para cada agente.

Los orígenes wittgensteinianos de la prioridad de las actitudes normativas

Esta idea de la prioridad de la actitud evaluativa por sobre el fundamento metafísico de la atribución de agencia e intenciones tiene sus orígenes en la filosofía de Wittgenstein. En un conocido parágrafo de *Investigaciones filosóficas* el autor afirma: “Mi actitud hacia él es una actitud hacia un alma. No soy de la opinión de que él tiene un alma” (IF, 2, IV). Más allá del sentido críptico que pueda tener este aforismo y la vaguedad del concepto de “alma”, la idea general refleja el espíritu de las posturas normativistas vistas anteriormente. La clave para interpretar esta frase de Wittgenstein se encuentra en el contraste entre actitud y opinión. El paradigma que se pone en cuestión aquí es el que da por sentado que una actitud debe tener siempre como fundamento una opinión. Tratar a un tercero como si tuviese un alma, es tratarlo de manera diferente a un mecanismo regido por leyes causales. Es decir, es reconocerlo como un agente racional cuyos comportamientos están basados en estados intencionales y reconocerlo como un sujeto de compromisos racionales y morales. Como explica Kripke (1982), Wittgenstein niega que las actitudes hacia otros que consideramos agentes no viene justificada por una creencia acerca de sus estados internos, como las intenciones o los deseos (p. 148).

Asumir una actitud valorativa como la que propone Bilgrami o Korsgaard no nos compromete con una tesis metafísica acerca de la relación causal entre ciertos estados mentales y la conducta manifiesta. Esto aparece muy claramente en los desarrollos de Wittgenstein, donde lo crucial no son las condiciones de verdad, por lo cual en estos planteos lo relevante no es señalar el estado de cosas que tornan verdadera la afirmación “S es un agente” o “S actuó con una intención ϕ”. Lo que es relevante es pensar en el juego de lenguaje propio de la atribución de agencia, intencionalidad y responsabilidad, es decir, las circunstancias en que realizamos esta atribución recíproca, las reglas que regulan esta práctica y el rol que esta juega en nuestro modo de vida. Lo que se aprende en el marco de este juego de lenguaje es ciertas actitudes reactivas valorativas hacia otros considerados agentes humanos y esta práctica no está fundamentada en creencias previas a la atribución misma.

Un aporte normativista desde la filosofía del lenguaje y de la mente

La sección anterior pretendió ofrecer una explicación del sentido que tiene la afirmación de que la agencia es esencialmente normativa. También intentó dejar en clara la tesis central de todo enfoque normativista sobre la intencionalidad, a saber, que la intención es un estado esencialmente normativo (y no un estado sujeto a presiones normativas). Por otro lado, se hizo hincapié en que este enfoque normativista no niega el rol causal de las intenciones en la producción de la acción y en el control de la conducta, sino que incluso propone un modelo para explicar su eficacia causal, modelo que integra la concepción normativa de los estados intencionales. Un rasgo compartido por los enfoques como los de Bilgrami y Korsgaard es el de tomar como fundamento el valor intrínseco de la agencia humana. Por esto mismo es que ambas posturas se enfrentan a la difícil tarea de fundamentar este valor último. Podríamos decir que a la base de estos enfoques se encuentran normas del ser, según la taxonomía desarrollada al comienzo.

Sin embargo las estrategias argumentativas detalladas anteriormente no son las únicas alternativas para defender una postura de la doble perspectiva sobre la acción. A continuación presentaremos un modo de desarrollar un enfoque normativista sobre la agencia y las intenciones que no toma como fundamento el carácter intrínsecamente valioso de la agencia humana. Nos centraremos puntualmente en la teoría sobre el contenido de Robert Brandom. Nuestra hipótesis aquí es doble. En primer lugar, consideramos que el enfoque de Brandom desafía la caracterización que presenta la literatura en la filosofía de la acción acerca de los enfoques de la doble perspectiva. En segundo lugar, los desarrollos de Brandom ofrecen un modelo de las intenciones que logra explicar en mayor detalle la tesis de que estas son estados normativos.

Durante la segunda mitad del siglo XX comenzó a extenderse dentro de la filosofía del lenguaje la idea de que el significado es esencialmente un fenómeno normativo. Esta tesis normativista no se vio limitada al caso de los significados lingüísticos sino que también fue defendida en relación al contenido mental en general. La idea básica en este contexto es que puede decirse que algo tiene contenido o significado si hay ciertas normas vigentes en juego. Esta postura admite una doble interpretación: o bien el contenido/significado es fuente de normatividad o bien la existencia de ciertas normas genera contenido/significado. Estas lecturas dan origen a dos

tesis normativistas diferentes (pero no excluyentes), a saber, la de la normatividad engendrada por el contenido/significado (*meaning/content engendered normativity*) y las del significado/contenido determinado por la normatividad (*meaning/content determining normativity*) (Glüer y Wikforss, 2009). Como señalan Glüer y Wikforss (2018), ambas posturas apelan por lo general a las normas de acción más que a las normas valorativas.

La pregunta que naturalmente surge en este punto es qué implicancias puede tener una concepción sobre los contenidos y significados para el tópico de la agencia y las intenciones. La respuesta breve alude a que los estados intencionales, entre ellos las intenciones, poseen contenido proposicional y por lo tanto el modo en que se conciba ese contenido tendrá relación directa en nuestra concepción de las intenciones. Así, si consideramos el aporte que la posición normativista en la filosofía del lenguaje y de la mente puede hacer a la filosofía de la acción, podemos analizar los alcances y posibilidades que pueda tener el proyecto de comprender los estados intencionales partiendo de una caracterización normativa de los mismos. Robert Brandom es uno de los autores más influyentes del normativismo en la actualidad. Su filosofía se centra en la reflexión sobre la naturaleza normativa del contenido conceptual pero también ha desarrollados las implicancias de esta perspectiva en relación a la acción. El pensamiento de Brandom es complejo y su obra es vasta y extensa, por lo cual no pretendemos hacer un desarrollo exhaustivo de sus ideas. Nuestro objetivo será mucho más modesto y se limitará a señalar aquellos elementos de su filosofía que fundamentan la tesis de que las intenciones son estados normativos.

Robert Brandom y el inferencialismo como propuesta para entender la tesis de las intenciones como compromisos

Uno de los pilares fundamentales de la filosofía de Brandom es la defensa de una postura inferencialista expresivista dentro de un enfoque pragmático normativista del contenido semántico en general. El inferencialismo, el expresivismo y la pragmática normativista arrojan las coordenadas en las que se inscribe su concepción de las intenciones y de la agencia, es por eso

que una breve caracterización de estas tesis servirá para situar el pensamiento de Brandom.

El inferencialismo y el expresivismo son dos aspectos de una postura anti-representacionalista. El inferencialismo rechaza la idea de que la noción de representación deba tomarse como primitiva y se propone como una tesis acerca de cómo se constituyen los contenidos semánticos. El contenido de un concepto se determina mediante el conjunto de los conceptos con los que está inferencialmente conectado (Brandom 2000). Por ejemplo, el concepto de "rojo" está inferencialmente conectado con los de "coloreado" y "espacialmente extendido". La correcta aplicación del concepto "rojo" habilita la aplicación de los otros conceptos. A su vez, prohíbe otras aplicaciones como las del concepto "verde" o "inmaterial". Para Brandom, lo propio de aquello que es conceptual es poder articularse de manera inferencial y lo que determina el contenido conceptual es el rol que un determinado ítem con contenido ocupa en esa red de inferencias[8].

El expresivismo, por su parte, es una postura anti-representacionalista en el sentido de que cuestiona la idea de que todas las expresiones significativas tengan como función representar aspectos de la realidad (Frápolli y Villanueva, 2013). Desde la perspectiva expresivista de Brandom, las nociones lógicas, semánticas (como las de verdad y referencia) y modales son medios de hacer explícitos compromisos del hablante y no tienen como función representar cómo es el mundo. El expresivismo en Brandom también responde a la idea de que aplicar conceptos es una forma de hacer explícito algo que está implícito en nuestras prácticas. Aquí la explicitud es identificada con la articulación conceptual. Aplicar conceptos es hacer explícito compromisos racionales que aprendemos de manera implícita en la práctica.

Tanto el expresivismo como el inferencialismo de Brandom se insertan dentro de una concepción pragmática normativa. El carácter pragmático de su postura reside en la idea de que entender un concepto es dominar su uso inferencial y que es el rol que cumplen en estas inferencias el que les otorga su contenido. El aspecto normativo de este pragmatismo está dado por los conceptos de compromiso y autorización: dominar un concepto es saber de manera implícita con cuales otros conceptos quedo comprometido *(committed)*, a que inferencias quedo autorizado (*entitled*) y cuáles quedan excluidas (*precluded*). Resumiendo su postura Brandom sostiene:

Pragmatism about the conceptual seeks to understand what it is explicitly to say or think that something is the case in terms of what one must implicitly know how (be able) to do. That the relevant sort of doing is a constellation of asserting and inferring, making claims and giving and asking for reasons for them, is the essence of rationalist or inferentialist pragmatism about the conceptual[9] (2000, p. 18).

Las tesis descriptas conciernen tanto a las expresiones lingüísticas como a las intenciones, ya que estas también tienen contenido. Así como los conceptos cuando están involucrados en creencias y sus afirmaciones explicitas representan compromisos discursivos (a creer/afirmar sus consecuencias inferenciales y a negar otras), el contenido conceptual de la intención la convierte en un compromiso práctico. Este compromiso práctico es el que el agente debe asumir con otras intenciones que están inferencialmente articuladas con su intención original y también es un compromiso con una serie de acciones que debe realizar para satisfacer su intención. Por recurrir a un ejemplo brindado por Brandom, si me comprometo a llevar a otra persona al aeropuerto mañana por la mañana, esta intención me compromete a formar otras intenciones tales como poner el despertador a la hora indicada y conducir hacia el lugar en donde se encuentre esa persona mañana a la mañana con suficiente tiempo como para llegar al aeropuerto a tiempo. A su vez, esta intención me inhabilita en sentido normativo a formar otras intenciones incompatibles, tales como dormir durante toda la mañana o planear otra actividad (Brandom, 2009). De esta manera Brandom entiende que una acción es intencional si es el resultado de un reconocimiento de un compromiso práctico por parte del agente que actúa. De esta manera puede apreciarse como la tesis inferencialista sobre el contenido se aplica a las intenciones y por qué estas son entendidas como compromisos (Brandom, 1994).

Brandom (1994) rescata la idea de Sellars que aprender a responder a una intención con una acción adecuada es parte de comprender el significado de la intención (en lo que respecta a su contenido) y de comprender el juego de lenguaje en el que se articulan las intenciones. Estos juegos de lenguaje están conformados por prácticas de atribución y auto-atribución de estados intencionales y es dentro del mismo juego que las intenciones adquieren un contenido. La perspectiva pragmática normativista explica cómo estas prácticas otorgan contenido y determinan

reglas de inferencia de manera implícita en prácticas sociales. Cuando aprendemos a dominar el juego de lenguaje de atribuir intenciones aprendemos las reglas implícitas de tal juego. Cuando un agente declara tener la intención de realizar ϕ, bajo la forma de "voy a hacer ϕ", no está expresando una creencia acerca de algo que ocurrirá en el futuro. Al auto-atribuirse una intención el agente está asumiendo un compromiso práctico: el de realizar una acción que la satisfaga y de formar otras intenciones inferencialmente articuladas con ϕ.

Este autor también se compromete a dar una respuesta al problema de la eficacia causal de las intenciones sobre la conducta. La dimensión normativa de la intención hace que especifique qué debe ser hecho. En este sentido, la intención puede producir una acción solo en la medida en que el agente esté entrenado en la práctica de responder de manera confiable a ese contenido. Así los estados deónticos como los estados del agente que posee determinada intención influyen en el orden causal por medio de las actitudes deónticas (1994). La actitud deóntica es la del reconocimiento del compromiso práctico en cuestión. En esta concepción de la intención las intenciones son causas, dado que en el agente entrenado correctamente el reconocimiento de compromisos prácticos produce comportamientos de manera confiable.

El concepto de agencia se articula con esta caracterización de la intención. Ser un agente es ser sujeto de un tipo distintivo de evaluación normativa, es decir, ser sujeto de obligaciones racionales, prohibiciones y permisos. Los agentes racionales son aquellos a los que se les puede demandar razones por lo que dicen, creen y hacen y se puede interpretar su comportamiento como el ejercicio de reconocimiento de compromisos prácticos. Esto significa ser una criatura que se encuentra en el "espacio de las razones" (*space of reasons*). La racionalidad práctica tiene que ver con dar razones del comportamiento y conocer los compromisos inferenciales a los que están sujetas mis intenciones. Este aspecto constitutivo de la agencia racional no es del todo independiente de las actitudes evaluativas. Si una criatura es indiferente a las razones e insensible a las relaciones de consecuencias racionales e incompatibilidades entre sus actitudes, metas y comportamientos, no hay motivos para tratarla como un agente racional.

Tratar a alguien como susceptible de que su actuar sea evaluado racionalmente es tratarlo como alguien de nosotros. Esta actitud normativa

es lo que Brandom denomina reconocimiento y adoptar esta actitud es reconocer un tipo de comunidad con quien es reconocido de esta forma. Esta actitud de reconocimiento implica tratar a otros como agentes responsables de su conducta. Explica Brandom:

> *The core of the notion of recognition is the idea that normative statuses such as rational responsibility and authority (commitment and entitlement) are fundamentally* social *statuses. They are in principle unintelligible apart from consideration of the practical attitudes of those who* hold *each other responsible,* acknowledge *each other's authority,* attribute *commitments and entitlements to one another*[10] (2009, p. 3).

Por eso es espacio de las razones es un espacio esencialmente social y la agencia no puede ser entendida fuera de este contexto público y normativo.

Consideraciones finales

En el presente capítulo hemos caracterizado las tesis generales que constituyen el núcleo de la teoría estándar de la acción. También hemos señalado como su paradigma causal basado en relaciones entre eventos se propone como la única alternativa filosófica viable para habilitar la investigación en psicología sobre las intenciones y la agencia humana. Entre las teorías rivales a las que se enfrenta la concepción estándar se encuentran los enfoques de doble perspectiva. La hipótesis general que hemos intentado defender es que la caracterización que se hace de estos enfoques no responde a la complejidad de sus planteos. El poco interés que estos han suscitado en el campo de la filosofía de la acción es la causa de que sus tesis se presenten de manera vaga y oscura y que se malinterprete sus objetivos.

Las consideraciones de Bilgrami y Korsgaard han servido para desarrollar la idea de que la agencia y los estados intencionales son esencialmente normativos. La vía que han escogido estos autores es la de poner como fundamento de sus enfoques cierto tipo de actitudes evaluativas. Las breves alusiones a la postura de Wittgenstein iluminada por Kripke han tenido como objetivo clarificar esta tesis de la prioridad de la actitud evaluativa por sobre la creencia metafísica. Más allá de los aspectos más específicos de cada enfoque la idea que se encuentra en la base parece

responder a una postura parece razonable y realista. Esta es la idea de que cuando identificamos a alguien como agente que actúa en relación a intenciones ya lo estamos considerando como sujeto de evaluaciones. Ese aspecto es independiente de nuestras opiniones metafísicas acerca de lo que ocurre en sus mentes y en las nuestras. De aquí no se sigue que no haya una descripción metafísica válida de lo que es una intención o de lo que es ser un agente, sino que tal descripción no puede tratar a la intención como mera causa eficiente dejando a un lado la dimensión normativa que es constitutiva. Un paradigma que se agote en el análisis causal y disposicionalista de la intención no logrará dar cuenta de su dimensión evaluativa.

Tras la exposición de las posturas de Bilgrami y Korsgaard hemos señalado que sus propuestas no son la única vía para comprender la idea que está a la base de las teorías de la doble perspectiva. El enfoque normativista del contenido defendido por Brandom proporciona el marco dentro del cual la concepción de las intenciones como estados normativos adquiere un mayor desarrollo. En este sentido, consideramos que un enfoque como el de Brandom posee ventajas explicativas considerables frente a la teoría estándar de la acción y frente a los demás enfoques de la doble perspectiva. Estas ventajas están dadas por el hecho de que la teoría inferencialista pragmática y normativista logra explicar por qué las intenciones están sujetas a ciertos principios normativos (aspecto reconocido por los mismos causalistas) y a su vez explica por qué deben ser consideradas estados normativos en base al hecho de que poseen contenido (aspecto reconocido tanto por los causalistas como por los defensores de la dualidad de perspectivas).

La exposición de las ideas centrales de algunas de las posturas de la doble perspectiva más influyentes ha cumplido con el objetivo de mostrar que no hay un desinterés en la relación causal entre la intención y la acción. Ninguno de los autores que defiende una postura normativista sobre la agencia niega esta relación. Inclusive en sus propuestas consideran un modelo que explica como el agente ejerce control sobre su conducta, modelo que tiene el concepto compromiso como elemento mediador entre un estado mental y una acción. En este sentido, consideramos que un paradigma de la doble perspectiva no inhabilita la investigación en psicología sobre la agencia. Inclusive este paradigma permite distinguir las

preguntas normativas de las preguntas empíricas. Por un lado están los compromisos prácticos racionales que un agente debe asumir dada una intención, compromisos que determinan qué acciones están en consonancia con la intención en cuestión. Pero por otro lado subsiste la cuestión acerca de qué mecanismos psicológicos están a la base del actuar en conformidad con una intención reconocida, la pregunta por la incidencia de factores como la motivación, las emociones o la información implícita proveniente del ambiente, etc., así como las pregunta acerca del correlato físico de tales mecanismos. Estas son preguntas empíricas cuya respuesta es sumamente relevante para comprender mejor nuestra agencia humana.

Para finalizar, cabe la aclaración de que este capítulo no se ha propuesto como una defensa de los enfoques de la doble perspectiva en el sentido de presentarlos como la explicación correcta acerca de la agencia humana y las intenciones. Esto implicaría analizar sus posibles objeciones y dificultades a las que se enfrentan y semejante tarea amerita un trabajo aparte. Nuestro objetivo ha sido simplemente ofrecer un desarrollo claro del significado de las tesis acerca de que la agencia es esencialmente normativa y de que las intenciones son estados normativos, así como ofrecer una explicación del vínculo causal entre la intención y la acción dentro del marco teórico de la doble perspectiva. Consideramos que a estos enfoques se les debería brindar una mayor atención en la literatura sobre la filosofía de la acción, ya que entre sus ventajas se encuentran las de reconocer el aspecto social de la agencia, el carácter público de las intenciones en tanto poseedoras de un contenido compartible y la dimensión temporalmente extendida de los agentes racionales humanos y su identidad narrativa. Además, desde este marco normativista se podría extender la explicación de las acciones corporales para dar cuenta también de la agencia mental, ya que nuestras acciones mentales también integran el espacio de las razones y son procesos de los cuales debemos dar cuenta y sobre los cuales tenemos responsabilidad. Pero la indagación en tales ventajas forma parte de una agenda de investigación a mediano y largo plazo y será objeto de futuros trabajos.

Referencias

Aguilar, J. y Buckareff, A. (Eds.), Causing Human Actions, New Perspectives on the Causal Theory of Action, pp. 183-198. Cambridge: The MIT press.

Audi, R. (1973). “Intending”. En *Journal of Philosophy*, 70, pp. 387-403.

Baer, J. (2008). Free Will Requires Determinism. En Baer, J. y Kaufman, J. y Baumeister, R. (Eds), *Are We Free? Psychology and Free Will*. New York: Oxford University Press.

Bargh, J., y Ferguson, M. (2000). Beyond Behaviorism: On the Automaticity of Higher Mental Processes. *Psychological Bulletin*, 126, pp. 925-45.

Bilgrami, A. (2006). *Self-Knowledge and Resentment*. Cambridge: Harvard University Press.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge MA: Harvard University Press.

—— (2000). *Articulating Reasons: An Introduction to Inferentialism.* Cambridge MA: Harvard University Press.

—— (2009). *Reason in Philosophy: Animating Ideas.* Cambridge MA: Harvard University Press.

—— (1984). Two Faces of Intention. En *Philosophical Review*, 93 (3), pp. 375-405.

—— (1987). *Intention, Plans, and Practical Reason*. Cambridge Mass; London: Harvard University Press.

Bratman, M. (2007). Structures of agency (2007). New York: Oxford University Press.

—— (2009). Intention, Practical Rationality, and Self-Governance. En *Ethics* 119 (3), pp. 411 443.

—— (2018). *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford: Oxford University Press.

Dancy, J. y Sandis, C. (Eds.). (2015). Philosophy of action: An anthology. Oxford: Wiley Blackwell.

Davidson, D. (1963). Actions, Reasons, and Causes. Reimpreso en Davidson (1980).

—— (1970). Mental events. Reimpreso en Davidson (1980).

—— (1980). *Essays on Actions and Events*, Oxford: Clarendon Press.

Enç, B. (2003). *How We Act: Causes, Reasons, and Intentions*. Oxford: Oxford University Press.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. En *Journal of Philosophy* 68, pp. 5-20.

Frápolli, N. y Villanueva, M. J. (2013). Frege, Sellars, Brandom: Expresivismo e inferencialismo semánticos. En Pérez Chico, D. (coord.), *Perspectivas en la filosofía del lenguaje*. Zaragoza: Prensas Universitarias de Zaragoza.

Goldman, A. (1970). *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Glüer, K, & Wikforss, A. (2009). Against Content Normativity. En *Mind*, 118, pp. 31-70.

Glüer, Kathrin & Wikforss, Åsa, “The Normativity of Meaning and Content”, The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (ed.), URL=https://plato.stanford.edu/archives/spr2018/entries/meaning-normativity/.

Howard, G. (2008). WhoseWill? How Free? En Baer, J. y Kaufman, J. y Baumeister, R. (Eds), *Are*

We Free? Psychology and Free Will. New York: Oxford University Press. New York: Oxford University Press.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Cambridge MA: Harvard University Press.

Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Mele, A. (2003). *Motivation and Agency*. Oxford: Oxford University Press.

—— (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.

—— (2010). Teleological Explanation of Actions: Anticausalism versus Causalism. En

Nichols, S. (2008). How Can Psychology Contribute to the FreeWill Debate? En Baer, J. y Kaufman, J. y Baumeister, R. (Eds), *Are We Free? Psychology and Free Will*. New York: Oxford University Press. New York: Oxford University Press.

Schlosser, M. (2011). Agency, Ownership, and the Standard Theory. En Aguilar, J., Buckareff, A. y Frankish, K. (Eds.), *New Waves in Philosophy of Action*, pp. 13-31. Basingstoke: Palgrave Macmillan.

—— "Agency", The Stanford Encyclopedia of Philosophy (Fall 2015 Edition), Edward N. Zalta (ed.), URL=https://plato.stanford.edu/archives/fall2015/entries/agency/.

Seager, W. (2000). Real Patterns and Surface Metaphysics. En Thompson, D., Ross, D. y Brook, A. (Ed.), *Dennett's philosophy, a comprehensive assessment*, pp. 95-129. Cambridge, MA y London, UK: MIT Press.

Searle, J. (1991). Response: Perception and the Satisfaction of Intentionality. En *John Searle and his Critics*, E. Lepore y R. Van Gulick (eds.), Oxford: Blackwell, pp. 181-192. Searle 1963.

Setiya, K. (2007). *Reasons without Rationalism*. Princeton: Princeton University Press.

Setiya, Kieran, "Intention", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), forthcoming URL=https://plato.stanford.edu/archives/fall2018/entries/intention/.Strawson, P. (1974). Freedom and Resentment and Other Essays. New York: Routledge.

Wilson, George and Shpall, Samuel, "Action", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL=https://plato.stanford.edu/archives/win2016/entries/action/.

Wittgenstein, L. (1953 [2017]). *Investigaciones filosóficas.* García Suárez, A. y Moulines, U. (Trad.), prólogo de Reguera Pérez. Madrid: Gredos.

---

[1] Desde la popularización de los escritos de Davidson en los 70, la teoría estándar de la acción

concibe las intenciones como estados mentales. En un comienzo primaba un enfoque reduccionista, inspirado en los primeros escritos davidsonianos, según el cual las intenciones están constituidas por deseos y creencias instrumentales (Goldman, 1970; Audi, 1973). Pero más adelante comenzó a popularizarse el modelo no-reduccionista de las intenciones. Esta perspectiva reconoce a las intenciones como estados mentales distintivos cuya naturaleza es diferente e irreductible a la de los deseos y creencias (Bratman, 1984; Mele, 2003). Este cambio de perspectiva también encuentra su expresión en los ensayos posteriores de Davidson.

[2] Esta estrategia de fundamentación encierra una serie de problemáticas y ambigüedades relativas a que no hay un sentido unívoco de "naturalismo" y al hecho de que adherir al naturalismo es, de por sí, algo intrínsecamente positivo, que no necesita ser justificado. Estas dificultades que atañen a todas las ramas de la filosofía que proponen una "naturalización" de su objeto de estudio, son resumidas irónicamente por William Seager (2000) cuando afirma "*Naturalism is supposed to be a Good Thing. So good in fact that everybody wants to be a naturalist, no matter what their views might be*" (Se supone que el naturalismo es algo bueno. Tan bueno que todos quieren ser naturalistas, sin importar cuáles sean sus puntos de vista) (p. 95).

[3] "…es que dentro de nuestros propios valores y evaluaciones podemos encontrar razones para no cometer suicidio de forma normativa y racional".

[4] Ellas tendrán que tomar la forma tal que no es que el compromiso de X lo llevó a hacer Y, sino que tal compromiso produjo que X tratase de vivir comprometido a hacer X.

[5] "*Wanton*" es el término utilizado por Frankfurt (1971) para aludir a una criatura hipotética que carezca de voliciones de segundo orden.

[6] "La mayoría del tiempo, nuestras razones para actuar emergen desde nuestras identidades más locales y contingentes. Pero parte de la fuerza normativa de esas razones emerge desde nuestro valor como seres humanos que necesitan de tales identidades".

[7] "Dado que no puedes actuar sin razones y que tu humanidad es la fuente de tales razones, debes valorar tu propia humanidad si actúas en lo absoluto".

[8] Estas consideraciones traen aparejadas dos consecuencias importantes de la semántica inferencialista de Brandom. La primera de ellas es la imposibilidad de poseer un concepto sin poseer, a su vez, muchos otros. La segunda es que la forma fundamental de lo conceptual es lo proposicional. Las proposiciones son las que pueden articularse de manera inferencial y no los conceptos considerados de manera aislada. Los conceptos se articulan en aserciones, creencias y pensamientos con contenido proposicional (obtenido de su rol en una red de inferencias) y es en esta articulación que los conceptos adquieren su contenido (Brandom, 2000, p. 12).

[9] El pragmatismo sobre lo conceptual busca entender qué es explicitamente decir o pensar que algo es el caso en términos de lo que uno debe implícitamente saber para hacerlo. El tipo relevante de "hacer" es una constelación de aseveraciones e inferencias, asegurar y requerir razones para tales

aseveraciones; esta es la escencia del racionalista o inferencialista pragmático sobre lo conceptual.

[10] El núcleo de la noción de reconocimiento es la idea de que estatus normativos tales como la responsabilidad racional y la autonomía (compromiso o derecho a) son estatus fundamentalmente sociales. Ellos son, en principio, ininteligibles aparte de la consideración de las actitudes prácticas de aquellos a quienes se les *toma* como responsables, *reconocer* la autoridad de otro, *atribuir* compromisos y derechos los unos a los otros.

ENFOQUES PARA EL ESTUDIO DE LO MENTAL

CAPÍTULO VII

Ciencia cognitiva como multidisciplina

Francisco Osorio

Introducción

Este capítulo realiza una breve discusión sobre qué sería lo multidisciplinario, específicamente en relación a la ciencia cognitiva. El concepto de multidisciplina está muy presente en las discusiones sobre la misión de las ciencias sociales o sobre cómo debería ser, pero pocas veces hay propuestas concretas de qué implica que este concepto o modo de trabajo entre las disciplinas científicas. De los posibles modelos a considerar, este texto se concentra en William Rehg (2000), Barbara Von Eckardt (2001), junto a Robert Goldstone y Loet Leydesdorff (2006).

El Modelo de los Estudios Científicos Críticos

El filósofo William Rehg se pregunta cómo construir el campo interdisciplinario de los estudios científicos críticos en su artículo “Critical science studies as argumentation theory: Who’s afraid of SSK?”, del año 2000. En palabras de Regh, este campo agrupa la tradición de la Escuela de Frankfurt (Horkheimer, Habermas), la teoría postcolonial y feminista (Harding), la epistemología social (Fuller, Longino), la crítica posmoderna del poder (Rouse) y los estudios de ciencia y tecnología (Brante).

En una definición general, según Rehg, los estudios científicos críticos se interesan en la ciencia empírica como un objeto de crítica. Atendiendo a la manera de definir el conocimiento científico, básicamente se pueden encontrar dos grandes tradiciones en este campo. Por un lado la Escuela Crítica (o de Frankfurt) y por otro la Escuela de Edimburgo, liderada por

David Bloor y Barry Barnes, que propone la Sociología del Conocimiento Científico (SCC, pues en el idioma inglés es SSK, Sociology of Scientific Knowledge), la cual examina el rol de los intereses sociales y los valores en la ciencia. Desde un punto de vista epistemológico, la SCC se caracteriza por un relativismo basado en el "ateismo justificacional", en palabras de Rehg, dado que enfáticamente niega que los argumentos puedan tener una fuerza intrínseca propia, o una fuerza que pueda ser distinguida del contexto social. En otras palabras, cualquier fuerza argumentativa que tengan la evidencia y las razones científicas, son prestadas de –y varían con– las circunstancias sociales contingentes. La SCC no postula que los factores sociales ayudan a arribar a posiciones justificadas, sino que tales factores son "finalmente decisivos" para la manera en que los científicos llegan a sus conclusiones.

De manera opuesta, la Escuela Crítica postula que si los científicos son sujetos racionales, debemos suponer que están motivados exclusivamente por la "fuerza del mejor argumento" en ausencia de evidencia en contrario. En otras palabras, las razones pueden tener una fuerza justificatoria intrínseca.

El problema que se le presenta a Rehg es que si desea proponer el campo de los Estudios Científicos Críticos, entonces debe encontrar una solución a las diferente maneras en que se entiende la argumentación, para así agrupar a filósofos, sociólogos críticos, historiados, antropólogos, feministas y "posmodernos". Rehg postula que este campo excluye a filósofos racionalistas que relegan la sociología a la génesis del error (Laudan) y a sociólogos relativistas anti-filosóficos y anti-normativos (Bloor, Collins, Woolgar).

La solución que propone Rehg es suscribirse a la Teoría de la Argumentación de F. H. Eemeren (1996) propuesta en "Fundamentals of Argumentation Theory". Para ello, Rehg sugiere un paso previo metodológico: adoptar el equivalente epistemológico del método de evitación del filósofo político John Rawls. Así, según Rehg, el campo interdisciplinario de los Estudios Científicos Críticos debe tratar de evitar las preguntas filosóficas intratables surgidas del debate de la argumentación racional. En los Estudios Científicos Críticos no se toma posición en absoluto sobre las preguntas epistemológicas más profundas de si los buenos argumentos tienen méritos intrínsicos o, de lo contrario, son siempre

perspectivas socioculturales. En cuanto teoría de la argumentación, los Estudios Científicos Críticos escrutinan los actuales procesos de argumentación. Es decir, la influencia de los factores sociales es importante solo si podemos identificar tales fuerzas y cómo ellas influencian una determinada controversia y su término. Por un lado, toma la argumentación en forma seria y reconoce la fuerza de los argumentos en tanto (como participantes) los reconocen. Por otro lado, no atribuye una fuerza intrínseca a los argumentos. Así, estudios de casos que fallen en descubrir factores sociales de influencia significativa, no fuerzan a los ateistas justificacionales a dejar sus creencias, como tampoco es el caso que estudios que identifiquen factores sociales significativos, vayan en contra de los argumentos racionales normativos. En resumen, según Rehg, ponemos a un lado –temporalmente– nuestras preguntas epistemológicas para los propósitos de la crítica interdisciplinaria.

Visto todo la anterior, William Rehg propone que los Estudios Científicos Críticos tienen como dominio el análisis empírico de las dimensiones socioinstitucionales de la argumentación científica. Un análisis dialéctico de la argumentación científica tiene los siguientes dos estándares o características:

- Carácter presuntivo: La argumentación científica descansa en una variedad de supuestos. Tales supuestos son considerados verdaderos hasta que no surgen dudas. Entre esos supuestos, es posible encontrar valores culturales que podría determinar lo que cuenta como evidencia para una hipótesis.
- Relevancia: Al alcanzar un consenso en ciencia, los participantes deben presumir que ninguno de los factores que influencian el consenso podrían, si salen a la luz, socavar la confianza en el consenso, dado sus objetivos de investigación (como sus participantes los definan).

De lo anterior, Rehg propone tres consecuencias. Primero, la irrelevancia debe ser distinguida de los efectos que puedan surgir del descubrimiento de supuestos particulares que sean falsos. La irrelevancia no es falsedad. Segundo, estos dos estándares no imponen un objetivo particular a la ciencia, dado que los participantes los definen. Tercero, los practicantes de

la argumentación crítica, al acercarse a un caso de controversia o de formación de consenso en ciencia, deben adoptar la perspectiva de los participantes reflejadas en sus creencias científicas o consensos. Estas creencias son objeto de un consenso general o, en caso de controversia, el consenso de una parte de la comunidad científica. El método que propone Rehg para solucionar las controversias es tomar conciencia de las posiciones de los consensos y los diferentes factores que constituyen su base motivacional. Al hacer su práctica explícita, los participantes operan con la sospecha de que factores sociales podrían ser significativos, o incluso determinantes, en el consenso. Hasta aquí, entonces, el desarrollo del modelo propuesto para el campo de los Estudios Científicos Críticos por Rehg.

Analizando su propuesta, se puede observar que para el caso de los Estudios Científicos Críticos, las diferentes disciplinas que entran al campo mantienen sus creencias, siendo excluidos solamente aquellos que se nieguen a participar de un campo interdisciplinario. Este carácter "respetuoso" de las diferencias crea la posibilidad de colaboración para un dominio. Dicho de otra manera, este campo es radicalmente interdisciplinario, porque nadie esta excluido por definición. De hecho, desde posmodernos hasta positivistas lógicos pueden formar el campo, si aceptan el dominio y los criterios de relevancia y de carácter presuntivo de los Estudios Científicos Críticos. En otras palabras, la controversia aquí es que no hay controversia. La solución es que "respetuosamente" todas estas tradiciones entran a participar del dominio.

En tanto solución para resolver los problemas que crea la formación de un campo interdisciplinario, no puedo sino reconocerla. Ahora bien, no considero que sea razonable tomar este camino, dado que crea más problemas que los que pretende resolver. Mi argumento es el siguiente: La especificación del dominio requiere la formación de un consenso en el campo interdisciplinario. Si esto es verdadero, entonces el criterio de relevancia se satisface por la mantención del consenso en función del dominio. El punto es que, en principio, la creación del consenso es posible, pero su mecanismo de mantención o de resolución de conflictos se basa en un supuesto difícil de sostener, que denominaré "terapia psicológica para conflictos de consenso". Según el campo de los Estudios Científicos Críticos, la toma de conciencia de las diferencias entre los problemas de los

investigadores causa la resolución de estos, dado que queda consciente un problema que antes era inconsciente. Sin embargo, no existe un mecanismo de resolución de conflictos, sino que un acrecentamiento de problemas y diferencias que nos exige decir “bueno, aun cuando tenemos estos problemas, sigamos juntos porque a todos nos interesa el domino”. La mantención del dominio a como de lugar no crea un campo justificado.

Sin embargo, es interesante rescatar algunas ideas de este modelo. En primer lugar, la formación de consenso no esta dada por los supuestos sustantivos compartidos por las disciplinas que concurren al campo, sino por la mantención del dominio. Es el dominio, entonces, el que genera el consenso. En segundo lugar, es interesante la idea que no se puede exigir a las disciplinas que abandonen sus supuestos, sino que pensar en construir un campo interdisciplinario desde supuestos que son, a veces, incompatibles. Tercero, la respuesta de Rehg es fundar la interdisciplinariedad en base a la Teoría de la Argumentación. Esta teoría, al igual que el pensamiento de Von Eckardt, considera los desarrollos de la filosofía de la ciencia desde Kuhn, sin desconocer los aportes de la tradición del positivismo lógico.

El Modelo de la Ciencia Cognitiva

La filósofa Bárbara Von Eckardt (2001) entrega un interesante argumento para analizar la multidisciplinariedad. Propone un modelo para fundamentar la ciencia cognitiva como un programa multidisciplinario, que denomina “marco de investigación”. Von Eckardt señala que la ciencia cognitiva fue fundada porque investigadores de diferentes disciplinas (inicialmente lingüística, inteligencia artificial y psicología, agregándose más tarde la filosofía, antropología, neurociencia y educación) se dieron cuenta, por un lado, que estaban interesados en estudiar básicamente el mismo conjunto de fenómenos (esto es, la inteligencia o cognición) usando un conjunto similar de constructos teóricos (esto es, las nociones de representación y computación) y, por otro lado, con el fin de estudiar seriamente el dominio de inteligencia/cognición, los puntos de vista y métodos de cada una de las disciplinas era requerido.

Del análisis sistemático de los artículos de la revista *Cognitive Science*,

Von Eckardt señala dos concepciones de la noción de multidisciplinariedad.

- La concepción localista. Esta concepción postula que un campo es multidisciplinario si el esfuerzo de investigación individual de sus científicos es, típicamente, multidisciplinario.
- La concepción holista. Esta concepción postula que un campo es multidisciplinario si es característico del campo que múltiples disciplinas contribuyan a la ejecución de su programa de investigación.

Si bien Von Eckardt señala que ambas nociones son abrasadas por los investigadores en ciencias cognitivas, ella postula que hay mejores argumentos para considerar la concepción holista de la multidisciplinariedad. La concepción localista postula que el foco de atención de lo que define una multidisciplina es el esfuerzo de investigación individual que se basa en la siguiente característica: involucra dos o más científicos de diferentes disciplinas, o métodos, o teorías, o referencias bibliográficas. Es decir, investigación individual quiere decir que una investigación particular califica como perteneciente al campo multidisciplinario de la ciencia cognitiva si combina dos o más elementos de esas disciplinas. El criterio es claro: la investigación que no involucre dos o más disciplinas no califica para formar el campo de la ciencia cognitiva.

Sumado a lo anterior, los científicos que adhieren a la concepción localista plantean que la ciencia cognitiva no solo es de hecho multidisciplinaria, sino que es esencialmente multidisciplinaria. Esto significa que la ciencia cognitiva no es psicología o inteligencia artificial, sino que es otra cosa, que es distinta a las ciencias especiales que concurren a su formación. Es, por decirlo de otra manera, un campo autónomo, una ciencia por sí misma.

La concepción holista que postula Von Eckardt pone el acento en la mantención del programa de investigación, lo cual permite que investigaciones unidisciplinarias califiquen como apropiadas a la ciencia cognitiva. Es decir, lo importante a mantener es el foco en el campo, lo que permitiría que la investigación unidisciplinaria sea parte de la ciencia cognitiva, en tanto ella se enfoque en el campo. Más específicamente, el programa de investigación de un campo es un conjunto de metas con

diferentes niveles y complejidades dirigido al propósito principal de responder las preguntas básicas del marco de investigación del campo dentro de las limitaciones establecidas por el dominio, los supuestos substantivos y los supuestos metodológicos de ese marco de investigación.

En palabras de Von Eckardt, lo que caracteriza a la concepción holista es el esfuerzo multidisciplinario, más que la investigación multidisciplinaria. Esto no quiere decir que la investigación multidisciplinaria no sea necesaria, todo lo contrario. Lo que dice es que, dado que históricamente los diferentes aspectos del programa de la ciencia cognitiva han sido la provincia de diferentes disciplinas, entonces es más racional pensar que la ciencia cognitiva logrará desarrollar su programa de investigación si cada disciplina hace su parte (participa de este esfuerzo), además de continuar con la investigación multidisciplinaria o interdisciplinaria.

¿Qué es una investigación interdisciplinaria? Von Eckardt ofrece un criterio para saber cuándo estamos en presencia de investigación interdisciplinaria: la concurrencia de dos disciplinas que forman un rótulo estándar reconocido por la comunidad científica y que se manifiesta, muchas veces, en el título de una revista científica. Por ejemplo, la neuropsicología es un rótulo estándar que está asociada a la revista *Neuropsychology Review*. Otro ejemplo es la antropología educacional, que por ejemplo se desarrolla en la revista *Anthropology y Education Quaterly*.

La mini Teoría del Marco de Investigación

La pregunta, entonces, es ¿qué significa marco de investigación? En este punto, el artículo de Von Eckardt remite al libro que ella escribió en 1993 (*What is Cognitive Science*?), donde se desarrolla en forma extensa su propuesta. Por lo tanto, abandonaremos el artículo por ahora y nos concentraremos en los argumentos que caracterizan a la ciencia cognitiva como marco de investigación multidisciplinario.

Von Eckardt (1993) señala que la ciencia cognitiva es un hecho sociológico establecido. Por este concepto, ella entiende las siguientes características que concurren a su institucionalización: (a) existe una sociedad de ciencia cognitiva; (b) se realizan congresos en forma regular; (c) existe una revista llamada ciencia cognitiva; (d) algunas editoriales de

prestigio tienen series de ciencia cognitiva; (e) existen magísteres, doctorados e incluso menciones en ciencia cognitiva de algunos pregrados; (f) centros especializados en investigación en ciencia cognitiva; y (g) mucho dinero se invierte en investigación en ciencia cognitiva.

Ella sostiene la existencia de una comunidad de investigadores en ciencia cognitiva. Esta comunidad esta guiada por un Esquema de Compromisos Compartidos (ECC) (*Framework of Shared Commitments* FSC). La mini Teoría de los Marcos de Investigación (MI) (*Research Frameworks* RF), funciona como un ECC para las comunidades científicas de la ciencia cognitiva. Por ECC, Von Eckardt entiende la versión de Kuhn de cómo funciona la ciencia, aunque con algunas modificaciones. Es decir, (a) las actividades científicas son típicamente llevadas a cabo por comunidades científicas, (b) estas actividades son de dos tipos. En lenguaje de Von Eckardt, (b1) las actividades que ocurren dentro del marco de compromisos compartidos (ciencia normal en lenguaje de Kuhn) y (b2) las que ocurren fuera del ECC, donde la comunidad opta por mantener su ECC o usar uno alternativo (revolución científica en lenguaje de Kuhn). Un ECC, entonces, guía o constriñe la actividad científica que ocurre a su interior. La pregunta es entonces, ¿cómo sabemos que estamos en presencia de un ECC? La repuesta, según Von Eckardt, es el uso de un marco de investigación (MI).

En resumen, Von Eckardt se encuentra con el problema que la ciencia cognitiva no posee las características de las ciencias maduras como la física (de hecho en el interior de la ciencia cognitiva solo concurren ciencias inmaduras como por ejemplo la psicología, la neurociencia y la antropología). Tampoco poseen sus explicaciones el grado exigido para ser consideradas explicaciones científicas (no pasarían el test que imaginariamente aplicaría Carl Hempel a la obra bajo análisis de Von Eckardt). Además, dado que ella considera correcto el análisis de la ciencia de Kuhn, se encuentra con el problema que la ciencia cognitiva tampoco pasaría el examen que imaginariamente haría Kuhn, pues en rigor las comunidades científicas que Kuhn fundamenta son las de las ciencias maduras. ¿Qué hacer entonces? Ella propone una mini teoría que explicaría la existencia de comunidades científicas en ciencia cognitiva realizando explicaciones científicas, si se acepta la siguiente tesis:

Tesis 1: En ciencias inmaduras, MI funciona como ECC para comunidades científicas.

Según Von Eckardt, un marco de investigación MI consiste en cuatro conjuntos de elementos: un conjunto D de supuestos que proveen una especificación preteorética del dominio bajo estudio, un conjunto P de preguntas básicas de investigación empírica formuladas preteoréticamente, un conjunto SS de supuestos substantivos que encarnan el enfoque tomado en la respuesta de las preguntas básicas y que constriñen las posibles respuestas a esas preguntas y, un conjunto SM de supuestos metodológicos. Por lo tanto:

$$MI = \langle D, P, SS, SM \rangle$$

Preteorético significa un conjunto se supuestos para conducir investigación, que no significa que sea una práctica dudosa por su carácter a priori o que defina un enfoque único de investigación, pase lo que pase. En vez de ello, es adoptar un conjunto de supuestos como una hipótesis general de trabajo acerca del carácter del dominio, que siempre puede ser abandonada.

(D) Una especificación preteorética de dominio incluye tres diferentes tipos de supuestos. (1) Un supuesto de identificación, (2) un supuesto de propiedades, y (3) un supuesto de agrupamiento. El primero sostiene que el dominio del marco de investigación consiste de fenómenos de tal y tal tipo. El segundo sostiene que el fenómeno seleccionado por el punto anterior tiene tales y tales propiedades. El tercero sostiene que los fenómenos seleccionados forman un conjunto teóricamente coherente de fenómenos o sistema.

La función de estos supuestos es “fijar” un dominio particular para buscar información (*domain of inquiry*). (Por dominio Von Eckardt entiende un conjunto de fenómenos acerca de los cuales existen preguntas científicas y respuestas). Estos supuestos son de “bajo nivel” en el sentido que el fenómeno puede ser considerado que existe con relativa independencia del aparato teórico o conceptual que uno adopte en orden a responder las preguntas acerca de ese fenómeno.

(P) Estas son preguntas acerca del dominio que, de una manera, dan a la investigación asociada con el marco su dirección. Estas preguntas básicas deben ser distinguidas de las preguntas empíricas derivadas. Estas últimas surgen, señala Von Eckardt, cuando añadimos los supuestos substantivos o cuando los datos han sido recolectados y las hipótesis específicas formuladas.

Von Eckardt postula que en el caso de la ciencia cognitiva, no es conveniente usar dos de las restricciones que la filosofía de la ciencia postula para la explicación científica: (1) que las preguntas básicas son preguntas por qué, y (2) que las preguntas básicas son todas preguntas acerca de los datos. Con respecto a la primera restricción, que Von Eckardt no está dispuesta a aceptar, ella señala que la ciencia cognitiva no posee esas preguntas, sino que preguntas qué y cómo (lo cual no la hace una forma errónea de empresa científica por tener diferentes preguntas). Con respecto a la segunda restricción, ella no esta dispuesta a aceptar la tesis que la actividad científica es principalmente el acrecentamiento de la información, ya sea por observación o experimentos controlados, con el propósito de avanzar en el conocimiento, donde las preguntas básicas están solo orientadas a los datos. Más bien, dice Von Eckardt, el principal objetivo de las preguntas básicas es responder las preguntas con respecto a los fenómenos que la ciencia investiga. Responder preguntas acerca de los datos es una actividad secundaria.

(SS) Los supuestos substantivos son proposiciones, algunas de las cuales definen o describen un sistema idealizado (estas proposiciones no son idénticas al sistema idealizado). Estas proposiciones no especifican el sistema idealizado en ningún detalle, simplemente bosquejan, a menudo en trazos gruesos, un tipo de sistema, específicamente el tipo relevante al proyecto de investigación a mano. Más específicamente, los supuestos substantivos caen en dos grupos: (1) supuestos de anclaje, y (2) supuestos de sistema. Los primeros señalan que el dominio D constituye, o es sustancialmente parecido, a un sistema del tipo X. Los segundos señalan que el sistema de tipo X tiene tales y cuales características.

Entonces, D (dominio) identifica una clase de fenómenos como de interés. P (preguntas) especifica lo que se quiere saber acerca de esos fenómenos. SS (supuestos substantivos) indica cómo esas preguntas deben ser respondidas. En particular, continúa Von Eckardt, SS nos dicen qué tipo

de sistema se supone que es el fenómeno bajo dominio y, por ello, constriñen los tipos de respuestas a las preguntas empíricas básicas que son relevantes.

(SM) Se refiere a las maneras apropiadas de estudiar el dominio en cuestión. Pueden ser (1) compartidas por la mayoría de la comunidad científica, o (2) propia del marco de investigación. Esto no se refiere a las técnicas o métodos específicos de investigación, sino que a los supuestos fundamentales de hacer investigación, tal que el abandono de esos supuestos metodológicos implica el abandono del marco de investigación. Hasta aquí, entonces, el desarrollo de la Teoría del Marco de Investigación como está desarrollado en Von Eckardt 1993.

¿Es multidisciplinaria la Ciencia Cognitiva? El análisis de Goldstone y Leydesdorff

Siguiendo el camino trazado por Von Eckardt (al menos una parte de él), Goldstone y Leydesdorff (2006) toman la revista *Cognitive Science* y hacen un estudio de sus "exportaciones e importaciones". Esta interesante manera de ver el tema consiste en preguntarse cuánto de la revista "sale" y es citado por otras publicaciones periódicas, así como que "artículos" (como si fuesen productos internados al país), pasan por la "aduana" de la revista y, por lo tanto, son citados dentro de la revista misma. Una de las creencias que desean estudiar es que la revista es más que nada psicológica, es decir, que ha sido atrapada por los psicólogos, dado que ellos tienden a mandar sus trabajos ahí. En otras palabras, que el campo ya no es multidisciplinario, sino que dominado por una disciplina, en este caso, la psicología (aunque lo mismo se podría decir de otras ciencias concurrentes como la neurociencia o la ciencia de la computación).

La manera en que realizan su estudio no es sino parte de las tendencias en ciencias sociales de esta parte de comienzos de siglo: analizar estadísticamente la revista a través del índice *Web of Science* (más conocido como ISI o también Thompson reuters humanities/sciences index). Por cierto, no todas las revistas están ahí (algo que los autores saben), pero es la herramienta que tienen para construir sus bases de datos (que corresponden a lo publicado en el año 2004).

Su primer descubrimiento es que la revista es citada por más revistas que las que ella cita. Las revistas que usan *Cognitive Science* están más el campo de la psicología cognitiva y la ciencia computacional, junto a la neurociencia (más asociada a la psicología). Sin embargo, la revista cita más a los artículos en el campo de la educación, lingüística y filosofía.

En resumen, los datos muestran que la revista es muy usada por la psicología, pero que la revista *Cognitive Science* no se limita a esa ciencia en articular, sino que es interdisciplinaria. Por este último concepto, los autores quieren señalar que la revista permite conectar campos que, de otra manera si la revista no existiese, no serían conectados de manera tan eficiente. En otras palabras, interdisciplinario también significa jugar un rol de importancia en la conexión de campos concurrentes al objeto de estudio. En sus palabras, la ciencia cognitiva ha hecho bien en descomponer su objeto de estudio en múltiples dimensiones, pero también porque ha sabido conectarlas e interrelacionarlas.

Finalmente, los autores también quieren señalar que lo interdisciplinario depende del cristal con que se lo mire. Así, las conclusiones a las que podemos llegar son muy diferentes si hacemos un análisis de las importaciones o desde las exportaciones de este país que se llama Ciencia Cognitiva en relación a sus países del mundo globalizado.

Una definición en proceso

Lo que queda, siempre, es mantener la pregunta abierta y observar cómo se da la discusión y estar sensible a los cambios de tendencias en la ciencia en general y, en particular, en las ciencias sociales. La propuesta de Von Eckardt es, hasta donde se puede aseverar, una de las más completas según la revisión bibliográfica sobre el tema, pero ello rápidamente indica que los límites de las propuestas que se pueden evaluar son, precisamente, aquellas de las que se tiene conocimiento, lo cual por estos días es abrumador dada la enorme cantidad de información disponible y la limitada capacidad de procesamiento o la poca capacidad de leer todo lo necesario.

Empero, una visión propia de comienzos del siglo XXI es evaluar las ideas por el impacto que ellas producen. Esto se entiende como la cantidad de citas que determinada obra tiene en diferentes medios académicos, tales

como revistas indexadas, libros y otros soportes susceptibles de medir. La clave está en la palabra medir, lo que permite comparar números, tal como el h-index de Google Scholar. En el caso de Barbara Von Eckardt, su artículo más citado de acuerdo a Google Scholar no era “Multidisciplinarity and cognitive science” (24 citas), sino que “Lexical and conceptual representation in beginning and proficient bilinguals” (597 citas), aunque su libro *What is cognitive science*? era su segunda obra más citada (252 citas). ¿Qué nos dice lo anterior? Que de acuerdo a estos números no parece que la propuesta de Von Eckardt fuese ampliamente aceptada por la comunidad académica. Más bien su impacto, hasta el momento, ha sido modesto.

Lo anterior es una extraña situación que no estaba disponible en el siglo XIX o hasta muy avanzado el siglo XX: el que la influencia de las ideas pudiera ser mensurada por su impacto. En otras palabras, no estamos evaluando el argumento en sí, estamos infiriendo que la propuesta hasta el momento no ha sido ampliamente aceptada por la comunidad académica, dado su bajo impacto. Visto desde otro punto de vista, no parece erróneo evaluar las propuestas por su nivel de aceptación, pues precisamente una propuesta filosófica de cómo debería funcionar el campo multidisciplinario de la ciencia cogntivida persigue, precisamente, establecer un conjunto de criterios y reglas que norman la actividad científica. Si ello se cumple o no es un tema empírico también.

El punto es que tal vez el impacto no sea la forma de evaluar un argumento, es decir, aceptar una definición muy específica del verbo impactar: cantidad de citas bibliográficas dentro de un dominio particular de publicaciones. Ello quedó muy claro con toda la discusión de comienzos del siglo XXI sobre *rankings*, índices y sistemas de evaluación para una multitud de cosas: qué universidad era la mejor, qué académicos podían seguir teniendo trabajo en la universidad, en qué revistas publicar y, en este caso, qué propuestas académicas aceptar.

Pues bien, seguiemos entonces durante los próximos años conservando la pregunta de qué hace a la ciencia cognitiva que posea la característica de multidisciplinaria y qué consecuencia tiene todo lo anterior.

Referencias

Goldstone, R. y Leydesdorff, L. (2006). The import and export of *Cognitive Science*. *Cognitive Science* 30(6): 983-993.

Rehg, W. (2000). Critical science studies as argumentation theory: Who's afraid of SSK? *Philosophy of the Social Sciences* 30(1): 33-48.

Von Eckardt, B. (2001) Multidisciplinarity and cognitive science. *Cognitive Science 25*: 453-470.

—— (1993). *What is Cognitive Science?* Cambridge, Mass.: MIT Press.

CAPÍTULO VIII

Elementos para una aproximación científica al estudio de la conversación

Edmundo Kronmüller

Introducción

Cada conversación es una conversación única y a la vez es todas las conversaciones. Cada conversación es única porque es una interacción situada en un espacio y un tiempo particular, con interlocutores concretos, que tienen conocimiento y propósitos específicos. Pero también cada conversación es todas las demás conversaciones porque los interlocutores traen expectativas sobre cómo se desarrolla una conversación que han adquirido en una historia de interacciones conversacionales. Al situarnos como oyentes en una conversación única, lo que hacemos es interpretar los que nos dice un hablante específico, al cual atribuímos intenciones de comunicarnos algo. De manera análoga, cuando nos situamos como oyentes en una conversación que es todas las conversaciones, lo que hacemos es interpretar a un hablante genérico, pero del cual solo tenemos expectativas sobre su comportamiento comunicativo.

El objetivo de este escrito es exponer una aproximación científica a los fenómenos comunicativos que se despliegan durante una conversación. Me focalizaré en el análisis de dos piezas fundamentales de esta aproximación científica. En primer lugar, haré una distinción entre dos estructuras de conocimiento, una de ellas para codificar información relativa al hablante específico y otra para codificar información relativa al hablante genérico. En segundo lugar, distinguiré los procesos involucrados en el tratamiento de la
información contenida en estas dos estructuras de conocimiento. El plan del escrito es como sigue. Primero, expondré la visión dominante en los

estudios actuales de comunicación en contextos conversacionales. Desde esta visión, la interpretación del habla conversacional es caracterizada como una inferencia de las intenciones comunicativas de un hablante. Segundo, mostraré los criterios que se necesitan para pasar desde una estrategia normativa de explicación, una estrategia intencional, hacia una estratégia propiamente científico-descriptiva, una estrategia de diseño. La estrategia intencional apela a estados volitivos y epistémicos de los agentes que se comunican para explicar los fenómenos de interpretación que tienen lugar en una conversación. La segunda estrategia en cambio, la estrategia de diseño, explica los fenómenos conversacionales apelando a la descripción de los mecanismos cognitivos que procesan la información comunicacional. Tercero, me focalizaré en diferenciar dos estructuras de conocimiento que son necesarias para explicar en términos mecanisistas la comunicación conversacional: el *common ground*, que hace de cada conversación una conversación única, y las expectativas pragmáticas, que hacen de cada conversación una misma conversación. Junto con diferenciar estas estructuras de conocimiento, presentaré la evidencia empírica –a partir de estudios experimentales– que sustenta la propuesta que la información del *common ground* se integra a la interpretación a través de un proceso de monitoreo de la comprensión y que las expectativas pragmáticas, en cambio, influencian la interpretación a través de mecanismos inferenciales simples o heurísticas. Finalmente, expondré algunas reflexiones relacionadas con este intento de dar una explicación mecanisista a los procesos de interpretación del lenguaje en conversaciones.

Dado que en una conversación ocurren distintos tipos de eventos comunicativos, como expresar una opinión o hacer un chiste, es necesario focalizarse en uno en particular. Para el objetivo de este escrito, tomaré como caso de análisis la referencia conversacional, y en específico la referencia definida (esto es, una frase nominal constituida por un artículo definido más un sustantivo). Así, el fenómeno a examinar es la interpretación por parte de un oyente de una referencia hecha por un hablante. Esta elección no es arbitraria, pues como expondré, gran parte de los avances hacia una explicación científica de la interpretación conversacional han tenido lugar en el contexto del estudio de la referencia definida.

La referencia definida y la inferencia de intenciones comunicativas

Las palabras no refieren. Son los hablantes quienes refieren, y para ello ocupan palabras. Al referir, lo que un hablante hace es guiar la atención de los interlocutores a una entidad, ya sea esta un objeto, una persona o incluso una idea. La tarea de los interlocutores es reconocer cuál es la entidad que el hablante intenta traer a la atención conjunta. En la mayoría de los casos, el significado convencional de las palabras no es suficiente para resolver la referencia e identificar la entidad en cuestión (Strawson, 1950). A modo de ejemplo, la expresión “el auto negro”, puede ser utilizada para referir a muchas entidades distintas que son autos y de color negro. Inversamente, existe indeterminación en tanto una misma entidad puede ser referida utilizando distintas expresiones, tales como mi auto, el cacharro o el escarabajo. En este sentido, una referencia definida especifica un elemento de una clase, pero solo en tanto clase, es decir, *un type*. Sin embargo, lo que se debe reconocer para que la comunicación sea efectiva es el elemento particular, es decir, *el token*. Esta inherente ambigüedad de la referencia obliga a los interlocutores a “ir más allá” del uso convencional de las palabras para lograr reconocer el objeto que el hablante quiere traer a la atención conjunta. Tradicionalmente, este “ir más allá” del significado convencional es caracterizado como una inferencia de un tipo muy particular: una inferencia de estados mentales. Así, lo que los interlocutores deben inferir en una conversación son las intenciones comunicativas del hablante.

El origen de esta caracterización de la comunicación como un proceso de inferencia de intenciones puede ser rastreado hasta la distinción ya clásica entre un enunciado y una oración (Strawson 1950), que fue hecha hace ya más de medio siglo en el contexto de la filosofía analítica. Un enunciado, por un lado, es una oración dicha en un lugar particular, por un hablante específico, y con un objetivo definido. La oración, por otro, es el aspecto semántico y sintáctico de las palabras involucradas en ella, las que tienen un carácter convencional, es decir, no contextual. Más aún, esta distinción tiene consecuencias en la misma definición de lo que entendemos por “significado”. En concreto, permite diferenciar entre dos nociones de significado. Por una parte, los enunciados tienen un significado situado, un

significado-del-hablante, en contraposición con el significado-de-la-oración (Grice 1957). Una manera cotidiana de expresar esta diferencia es entre lo-que-se-dice, que corresponde a la oración, y lo-que-se-quiere-decir, que corresponde al enunciado (en inglés esta distinción se hace utilizando los verbos *to mean* versus *to say*). Ya que frente a un enunciado se debe reconocer lo que el hablante quiere decir con lo que está diciendo, tradicionalmente se ha entendido a la comunicación como un proceso de inferencia de intenciones.

La racionalidad en la comunicación: el principio de cooperación

Si aceptamos que la caracterización de la comunicación como un proceso inferencial (Grice, 1957, 1975), se hace del todo relevante indagar en este proceso de inferencia. Como toda inferencia, la inferencia de intenciones comunicativas requiere de información dada y de supuestos. En términos normativos, el significado del hablante es inferido considerando (a) la conducta verbal –y no verbal– del hablante, y en especial el uso convencional de las palabras que componen la oración enunciada; (b) información contextual tales como el lugar de la interacción, los estados epistémicos del hablante (lo que sabe y no sabe), el propósito particular de la conversación y el conocimiento general del mundo circundante; y (c) el supuesto que el hablante es un agente racional que trata de hacer una contribución útil y oportuna a la conversación. Focalicémonos en esto último.

Uno de los supuestos de racionalidad básicos que guían la conversación es que los interlocutores harán su contribución de la manera más clara posible, en el momento adecuado, y para los propósitos de la conversación. Debido a que apela a una actitud de los hablantes que tiende a la colaboración en la interacción de manera de hacerla más fluida, es que se ha denominado a este supuesto el principio de cooperación (Grice, 1975). Es en virtud de este principio que la comunicación durante una conversación es posible. Una especificación de este principio de cooperación es la expectativa que tendrían los interlocutores de que mutuamente intentarán proveerse de la información necesaria para garantizar la comprensión. Así,

evitarán entregar menos información de la necesaria, pues de lo contrario sería una contribución conversacional ambigua. Así también, los interlocutores se esforzarán por no dar más información de la necesaria, evitando así entregar información irrelevante o de poco uso para los efectos de la presente conversación. En este sentido, tradicionalmente se asume que la racionalidad en la comunicación apela a cierto sentido de optimización de la información: minimizar la probabilidad de malos entendidos entregando la cantidad óptima información para reconocer la entidad referida.

Para ejemplificar, la expectativa de informatividad óptima llevaría a un oyente a inferir que, ante el enunciado "mira el auto grande", el auto referido se encuentra entre otros autos más pequeños; esto es, que en el dominio referencial existen otros autos que son de menor tamaño que el que se quiere traer a la atención conjunta. Del mismo modo, ante el enunciado "el auto", el oyente puede inferir que no existe otro auto a la vista para el hablante. Si en el primer caso el hablante sobre-informara y especificara el tamaño del auto sin otro auto posible, el oyente se confundiría y buscaría alguna razón para explicar el exceso de información entregado. Podría, por ejemplo, pensar que el hablante tiene acceso visual a un auto más pequeño que él no puede ver (Sedivy, 2003; Kronmüller *et al.*, 2014). En el lado opuesto, si el hablante enunciara simplemente "el auto" frente a dos autos, sería una contribución aparentemente ambigua, lo que también llevaría al oyente a buscar alguna razón para tal ambigüedad (por ejemplo, el oyente puede pensar que el hablante no puede ver el otro auto que él sí ve porque está ocluido –desde la perspectiva visual del hablante– por un bus).

El hablante de los supuestos y el hablante de las intenciones

La concepción de la comunicación como una inferencia nos permite escindir a un hablante en dos. Uno es el hablante del cual es necesario inferir las intenciones comunicativas. Esto implica hacer inferencias sobre sus estados epistémicos y volitivos. Es en este sentido en el que se puede decir que este es un hablante específico, con nombre, cara y estados mentales particulares. El otro hablante, por el contrario, es uno genérico. De este solo se espera cierto comportamiento, pero no porque sea ese hablante, sino porque es un hablante: un miembro de la categoría de agentes que se

comportan de una determinada manera. Esta determinada manera es la cooperación.

La estrategia intencional y la estrategia de diseño en la explicación de la interpretación del lenguaje hablado durante una conversación: del agente al mecanismo

La visión de la comunicación como un proceso inferencial podría ser considerado un gran avance en el estudio del lenguaje. Esta visión enfatiza la importancia de la comunicación cara a cara como el lugar fundamental donde se desarrolla el lenguaje. Es indiscutible que este desarrollo y giro se produjo en el contexto de la discusión filosófica, y específicamente de la filosofía analítica (este es el caso para muchos de los problemas y conceptualizaciones que son tratados por disciplinas empíricas, y en especial la psicología; la conciencia y la percepción son casos paradigmáticos de esto). Desde la filosofía, la explicación de la comunicación en la conversación tiene un carácter normativo. Esto quiere decir que la propuesta no plantea cómo se lleva a cabo el proceso, sino cuáles son los elementos que deben estar presentes para que el proceso pueda, en principio, ser realizado. Citando a Strawson: “En asegurar este resultado [interpretar una referencia definida], el contexto de un enunciado es de una importancia imposible de exagerar; y por *contexto* digo, al menos, el tiempo, el lugar, la situación, la identidad del hablante, los sujetos que conforman el principal foco de interés y las historias personales de los oyentes a los cuales el hablante está interpelando” (Strawson, 1950, p. 336).

La estrategia intencional, la estrategia de diseño y la inferencia

Hasta ahora he expuesto una explicación de la comunicación que la caracteriza como una inferencia que hace un agente. Es una estrategia intencional de explicación (Dennett, 1971). La estrategia intencional explica el comportamiento de un agente mediante la atribución a este de estados volitivos y epistémicos. Junto con esta atribución de estados intencionales, esta estrategia asume que el agente actúa de manera racional. Esta racionalidad es entendida en su forma más básica, es decir, como una

organización de la acción de manera de lograr los objetivos del agente descritos en sus estados epistémicos y volitivos: es decir, el agente generará un curso de acción para alcanzar sus objetivos [definidos en términos de querer algo o desear algo] considerando sus estados epistémicos [sus creencias sobre estados del mundo]. Así, si un agente quiere que otro agente reconozca una entidad como referente de su acción de referir, este primer agente buscará el camino óptimo para que esto suceda, entregando el tipo y la cantidad de información que sea lo más fácil de procesar y de la forma más directa. No tendría sentido si frente a una manzana y un durazno, un hablante dijera "quiero una rama" o más sutil aun "la fruta que quiero comer proviene de un árbol cuyo nombre científico es *malus domestica*". Claramente, decir "quiero la manzana" es más óptimo en tanto requiere menos esfuerzo de producción por parte del hablante y menos esfuerzo de interpretación por parte del oyente; es más directo para lograr que el interlocutor reconozca a la manzana como la entidad referida. Esta racionalidad es mínima, en tanto su única característica es definir la acción para lograr el objetivo. Este supuesto de racionalidad es el que está detrás de las máximas conversionales que se derivan del principio de cooperación antes expuesto (Grice, 1975).

Sin embargo, si queremos dar una explicación "científica" a la comprensión del lenguaje en uso, debemos apelar a una estrategia que apunte a la descripción de los mecanismos causales que producen la interpretación del lenguaje conversacional. En términos ortodoxos, esta estrategia implica entender a un interlocutor como un sistema cuyo comportamiento se explica en términos del cómo funciona. Esta estrategia es la llamada estrategia de diseño (Dennett, 1971). El trabajo principal para este tipo de explicación es describir los mecanismos –en tanto estructuras y procesos– que producen el comportamiento. Importantemente, estos mecanismos se comportan de manera sistemática y regular, en tanto son parte del entramado causal del mundo natural. Es así, un intento por naturalizar la explicación de la comunicación.

Tomando en cuenta esta estrategia de explicación, que es la propia de las ciencias cognitivas y psicológicas, el aporte principal de una visión normativa de la comprensión del lenguaje, en particular la estrategia intencional, es la caracterización de la comunicación como un proceso inferencial. Este es el *link* que permite el paso de una estrategia a otra. Es

importante esta aclaración, pues la propuesta tradicional, que he expuesto arriba, ha sido modificada e incluso fuente de críticas fundadas y profundas (Sperber y Wilson, 1986; Levinson, 2000). Es importante puntualizar además que ninguna de las estrategias es mejor que la otra, son solo dos posibilidades de explicación. Las ciencias cognitivas y psicológicas, al menos su versión más ortodoxa, se mueven principalmente en la estrategia de diseño, en tanto esta define a la cognición como un cúmulo de mecanismos que procesan información.

Pasos para una explicación desde la estrategia de diseño

Para abordar de manera naturalista la interpretación del lenguaje hablado durante una conversación, es necesario describir los mecanismos cognitivos que la explican. Y para el estudio de la referencia definida, en particular, lo que se debe describir son los mecanismos que explican el reconocimiento, por parte del oyente, de cuál es el referente que el hablante intenta traer a la atención conjunta. Cinco son las dimensiones que deben entenderse, definirse y describirse para lograr este tipo de explicación.

En primer lugar se debe definir a la entidad que se quiere describir. Para el caso de las ciencias cognitivas –al menos de la mayoría de la investigación en ciencias cognitivas– se entiende que se está describiendo a un sistema que procesa información. Este es un sistema complejo, constituido por estados internos que, junto con la estimulación externa y el *feedback* de su propia acción, determinan su comportamiento.

En segundo lugar, se debe describir los mecanismos para procesar la información. En términos muy generales, un mecanismo se define como una estructura que ejecuta una determinada función en virtud de sus partes, la operación de esas partes y la organización “orquestada” de estas operaciones (Bechtel, 2008). Esta descripción es transversal a muchas de las disciplinas de la ciencia moderna, como la biología y la genética, y puede ser rastreada hasta los trabajos de Demócrito y Epicúreo. Con esta definición, la descripción de mecanismos implica describir las estructuras y su procesamiento. Para el caso de un sistema que procesa información, en concreto, se deben describir las estructuras de conocimiento y codificación de información y las maneras en que esta información es procesada y

transformada.

Tercero, se debe describir de manera muy detallada la tarea que el sistema debe ejecutar en el medio en que se encuentra. Por ejemplo, escapar de predadores, encontrar alimento o coordinar acciones con otros congéneres.

Cuarto, se deben describir las restricciones que tiene el sistema en la ejecución de la tarea que debe realizar, tanto del medio como de su propia organización. Entre las restricciones más comunes para un organismo se encuentran las de tiempo (las más de las veces se debe ejecutar una acción de manera rápida) y de capacidad de procesamiento de la información (en general existe abundante información pero limitada capacidad para procesarla).

Y quinto, es necesario hacer una descripción profunda del medio en el cual este sistema se encuentra. Esta es quizás una de las tareas menos exploradas por la ciencia cognitiva. Proyectos como los de Gibson (1979), Simon (1956) y en el último tiempo Gigerenzer (1999), son intentos por describir de manera profunda el ambiente en el que se desenvuelve un sistema cognitivo. Una de las metáforas que mejor expresa esto es la definición de la cognición como el trazo que va dejando una tijera al cortar, cuya primera hoja es el organismo (o sistema) y la segunda el ambiente (Simon, 1956).

Considerando estos pasos, en el siguiente apartado me focalizaré en la tarea de dar una explicación de la interpretación de la referencia definida en contextos conversacionales en términos del funcionamiento del sistema, o lo que es lo mismo, describir el diseño que permite este funcionamiento. El énfasis estará dado en la descripción de las estructuras de conocimiento relacionadas con el hablante y las maneras de transformar y procesar esta información para lograr inferir cuál es la entidad a la que el hablante está refiriendo.

La “infraestructura cognitiva” para la interpretación de la referencia definida en una conversación: procesamiento basado en la información de los hablantes específico y genérico

No es casualidad que el abordaje científico de la comunicación durante una

conversación cotidiana tenga como centro de gravitación el dilucidar las estructuras de conocimiento que codifican información sobre el hablante y el tipo de procesamiento de esta información. La importancia de describir los mecanismos asociados al procesamiento de la información del hablante –y no de la información asociada a la sintaxis de un enunciado, por ejemplo– está determinada por la propia conceptualización de la comunicación como un proceso de coordinación de perspectivas entre, e intenciones entre, interlocutores. Es también coherente con el hecho que la pragmática, que es el área de la lingüística que estudia el lenguaje situado y en uso, sea definida como la disciplina que estudia los fenómenos del lenguaje donde es necesario apelar al hablante para poder, precisamente, dar cuenta de estos fenómenos (Levinson, 1983; Senft, 2014).

Como he expuesto, a partir de la estrategia intencional de explicación se puede hacer una distinción que es del todo relevante para describir el diseño de los mecanismos que nos interesan. Esta distinción es del tipo de estructuras de conocimiento que codifican información sobre los interlocutores. Tomando la definición del significado-del-hablante en términos de intenciones comunicativas, es claro que parte importante de la información que se necesita del hablante es sobre sus estados mentales. Por supuesto, se debe inferir lo que el hablante quiere decir con lo que está diciendo, dado los estados epistémicos y volitivos de este. Imputar estados mentales a los otros es cosa cotidiana y es parte importante de lo que se necesita para poder desenvolverse en interacciones con otros; es parte de nuestra sociabilidad en tanto permite coordinar nuestras acciones. Esta información constituye una estructura de conocimiento que va actualizándose in-situ, pues se aplica a una persona en un contexto particular que, en el caso de la conversación, es el hablante específico del cual inferimos intenciones comunicativas. Esta estructura de conocimiento se ha denominado conocimiento mutuo (Clark y Marshall, 1981) y más específicamente *common ground* (Clark, 1996). Este conocimiento mutuo, contingente y particular a cada conversación, la hace una instancia comunicativa única.

Otra pieza importante de información tiene relación con la serie de supuestos que se necesitan para poder inferir las intenciones del otro. Estas expectativas pragmáticas apuntan a sistematizar cómo los hablantes en general se comportan en una conversación. Una de las descripciones más

acabadas de este tipo de expectativas, aplicado a la comunicación, es el supuesto que los hablantes cooperarán de manera de hacer de su contribución a la conversación una que sea adecuada e interpretable, ya expuesto más arriba. Estos supuestos tienen un carácter muy diferente a la información relativa a los estados mentales de los hablantes. Ciertamente, se puede suponer que el hablante se comportará de determinada manera en tanto es un hablante como todos los demás hablantes: un hablante genérico. No es necesario, en este caso, apelar a estados internos contingentes. A diferencia del *common ground*, la estructura de conocimiento que se constituye por estas expectativas es permanente y se consolida a través de todas las interacciones comunicativas que tenemos durante la vida. Es conocimiento sobre patrones generales de conducta comunicativa y por lo tanto hace que cada conversación sea todas las conversaciones previas y futuras.

El procesamiento de la información del hablante específico: *common ground* y proceso de monitoreo de la comprensión

El concepto de *common ground* se hace cargo de lo que se llamó la "paradoja del conocimiento mutuo". Para que la definición de significado del hablante se sostenga, es necesario apelar a un conocimiento mutuo (Schiffer, 1972). Sin embargo, un conocimiento de este tipo solo puede ser representado como un set infinito de proposiciones anidadas del tipo "yo sé que tú sabes que yo sé que tú sabes...". Esta estructura representacional infinita es de una complejidad intratable para un sistema que la requiere para inferir intenciones comunicativas en tiempo real durante una conversación –incluso puede argumentarse que es una estructura implausible como representación cognitiva–.

Frente a esto, H. H. Clark y Marshall (1981) postulan un mecanismo cognitivo viable para inferir conocimiento mutuo con más o menos certeza: las heurísticas de co-presencia. Mediante estas heurísticas, el conocimiento mutuo puede ser inferido basándose en distintos tipos de evidencia. El primer tipo es la co-presencia física, que refiere una situación en la cual la información es percibida por los interlocutores en tanto están presentes en el mismo lugar y al mismo tiempo. Si dos personas miran hacia un árbol,

puede inferirse que es conocimiento común que el árbol está ahí. El segundo tipo de evidencia es la co-presencia lingüística, que refiere a los casos en que la información en introducida en el discurso que se va construyendo en una conversación. Por ejemplo, si un objeto ha sido nombrado como “el árbol” durante una conversación, puede inferirse que es conocimiento mutuo que ese árbol ha sido nombrado como “el árbol” (a diferencia, por ejemplo, de “el manzano”). Finalmente, la evidencia más débil es la pertenencia a una misma comunidad, ya sea un país, una ciudad o un grupo. Por ejemplo, dos personas nativas de un mismo idioma pueden asumir que es conocimiento común el significado convencional de ciertas palabras. Así, mediante estas heurísticas se puede inferir conocimiento común, que toma la forma particular de *common ground*, en tanto es una base sobre la cual se comunican dos personas.

De la noción de *common ground* se derivan hipótesis claras sobre cómo la personas interpretan un enunciado. Dentro de ellas, la más fuerte postula que en una conversación los interlocutores deben hacer sus contribuciones y sus interpretaciones diseñadas para el interlocutor específico, que es la persona en la conversación con la cual se ha construido –o se asume– *common ground*. Existe extensa evidencia de que efectivamente los interlocutores producen e interpretan sus contribuciones a la conversación para el hablante específico. En la producción de enunciados, se ha mostrado que los hablantes acortan sus descripciones para referirse nuevamente a un objeto cuando esta repetición es con el mismo hablante, lo que sugiere que a medida que el *common ground* crece, menos información en el enunciado (en este caso en la descripción) es complementada con información en el *common ground* (Clark y Wilkes-Gibbs, 1986). Por ejemplo, una primera vez un hablante puede referirse a un árbol como “el árbol de manzanas”, pero a medida que continúa la conversación, se referirá a esa misma entidad como “el árbol”. En la interpretación de enunciados, Clark, Schreuder y Buttrick (1983) mostraron que un oyente está más dispuesto a aceptar un artículo demostrativo, como esta o esa, cuando el referente es saliente entre otros referentes, pero no cuando todos los potenciales referentes son similares en saliencia. Al preguntar a una persona sobre cómo describiría el color de “esta flor” al mostrarle una foto con muchas flores, los oyentes comenzarán inmediatamente a describir el color de una flor particularmente grande entre otras flores pequeñas, pero pedirán más especificaciones

cuando todas las flores son del mismo tamaño. La explicación para esta asimetría es que el oyente asume que hay una flor más grande que la otra y que tanto el hablante como el oyente saben eso, pues es *common ground*.

Esta evidencia, y mucha otra, muestra que efectivamente el *common ground* guía la interpretación y la producción de contribuciones conversacionales. La pregunta que cabe es por el procesamiento de esa información: ¿Cómo es que esta información es transformada de manera de ser integrada a la interpretación? Existe amplia evidencia que el *common ground* es parte de procesos de monitoreo de la comprensión, es decir, procesos cognitivos complejos que de manera exhaustiva van corrigiendo interpretaciones tentativas en caso de que exista alguna inconsistencia entre esta y el estado de las cosas, lo que puede, potencialmente, provocar malos entendidos.

Keysar y colaboradores (1998, 2000, 2003) encontraron que las personas corrigen las interpretaciones de una referencia definida apelando al common ground después de que una interpretación “egocéntrica” ha tenido lugar. Es decir, existe una interpretación que no considera la información del hablante específico y que es posteriormente corregida de manera que sea consistente con el *common ground* (a lo que el hablante efectivamente puede referir). Este perfil temporal de la interpretación, en la que se puede observar una interpretación egocéntrica que es posteriormente corregida por un proceso de monitoreo basado en *common ground*, es posible debido a la naturaleza secuencial que tiene el habla. Los oyentes hacen interpretaciones (elecciones) referenciales a medida que van recibiendo la información auditiva secuencialmente. Por ejemplo, si en un dominio referencial hay un camión y dos autos (uno rojo y otro negro), frente a la instrucción de un hablante “mira el auto negro” se puede observar la siguiente secuencia: a) al escuchar “mira el”, los oyentes no tienen cómo decidir qué objeto es el que el hablante quiere atraer a la atención; b) al escuchar “auto”, los oyentes pueden considerar a los dos autos como referentes y descartar al camión; c) al escuchar “negro”, los oyentes pueden finalmente reconocer el objeto al que el hablante se refiere.

Para observar la interpretación egocéntrica y el momento en que se gatilla el proceso de monitoreo, es necesario dividir el dominio referencial en dos: el compartido por los interlocutores y el privado del oyente. Por ejemplo, el auto rojo, el auto negro y el camión son parte del dominio

referencial compartido entre oyente y hablante. Ahora, si además hay otro auto negro dentro de una máquina de lavado que solo el oyente puede mirar, y este oyente sabe que el hablante no sabe de su existencia, entonces este auto negro es del dominio referencial privado del oyente. El hablante no podría referir a este segundo auto negro pues no sabe que existe. Por lo mismo, el oyente debería considerar al auto negro compartido como el referente de la instrucción "mira el auto negro". Sin embargo, Keysar *et al.,* (2000) mostraron que por breves momentos los oyentes sí consideran a objetos del dominio referencial privado como potenciales referentes, pero esta "mala" interpretación es rápidamente corregida por el proceso de monitoreo de la comprensión a través del *common ground*. Siguiendo el ejemplo, al escuchar "negro" el oyente podrá experienciar ambigüedad, pues no sabría a cuál de los dos autos el hablante se refiere, hasta que monitoreando la comprensión reconocerá el objeto que está en el dominio referencial compartido como el objeto que el hablante quiere traer a la atención conjunta.

El procesamiento de la información del hablante genérico: expectativas pragmáticas y heurísticas simples

Las expectativas pragmáticas nos guían para determinar cómo nuestros interlocutores se deberían comportar. Por ejemplo, esperamos que si hacemos una pregunta el otro nos contestará. Estas expectativas las sostenemos para todos los hablantes con los que interactuamos, no porque sean esos hablantes específicos, sino porque son hablantes en general: son miembros de la clase de agentes de los cuales cabe suponer cierta manera de comportarse en ciertas situaciones. Son llamadas expectativas pragmáticas porque son supuestos sobre la conducta de personas en contextos comunicativos. De manera análoga al *common ground* para el hablante específico, se puede pensar que estas expectativas son sobre un hablante genérico. Es decir, ellas configuran una estructura de conocimiento que contiene información sobre patrones generales de conducta de los hablantes, sin especificar conocimiento compartido ni ningún otro tipo de información específica del hablante. Podemos decir entonces que se constituyen por patrones regulares de conducta comunicativa que son

aprendidos a través de todas las conversaciones previas en las que el sujeto ha estado involucrado.

Un caso particular de expectativa pragmática para el caso de la referencia definida es la expectativa de consistencia: los oyentes esperan que los hablantes sean consistentes en la manera en que refieren a las cosas. Se ha observado que, durante una conversación, los interlocutores convergen en nombres y expresiones para los referentes que traen a la atención conjunta (Garrod y Anderson, 1987). Los hablantes toman decisiones sobre cómo nombrar los objetos, por ejemplo, nombrar un auto como el auto negro y no como el auto grande, mi auto o el auto viejo, y tienden a mantener estos nombres al referir nuevamente al mismo objeto. Complementariamente, los hablantes tienden a evitar el uso de una misma expresión para referirse a dos objetos distintos. Estas regularidades permiten a los oyentes inferir de manera simple si el referente es un objeto nuevo o un objeto ya nombrado con anterioridad. Cuando un oyente escucha una expresión utilizada previamente por el hablante, espera que este se esté refiriendo a un mismo objeto nombrado previamente con esa misma forma lingüística. Del mismo modo, cuando un hablante utiliza una nueva expresión que no ha sido utilizada previamente en el diálogo, el oyente tiene buenas razones para interpretar que el hablante se está refiriendo a un objeto nuevo (Metzing y Brennan, 2003; Kronmüller y Barr, 2007).

Inicialmente se propuso que esta expectativa de consistencia estaba ligada a un pacto conceptual sobre cómo nombrar y conceptualizar los posibles referentes (Brennan y Clark, 1996). Este pacto, debido a que es un acuerdo –implícito las más de las veces– con un hablante específico, es conocimiento compartido entre los interlocutores y por lo tanto solo se esperaría consistencia con los patrones de referencia establecidos con el hablante con quien se tiene el pacto, y no con otro hablante no presente en la interacción donde este se estableció y por lo tanto ignorante sobre su existencia. Entonces, un oyente tiene la expectativa que su interlocutor usará una expresión distinta para referirse a un referente distinto. Si el hablante utiliza la expresión el auto grande, el oyente está compelido a buscar un auto que no sea el auto negro, pues este último fue referido precisamente como el auto negro previamente (está conceptualizado como tal). Sin embargo, si un nuevo hablante (ignorante sobre la existencia de un

pacto conceptual) utiliza la expresión “el auto grande”, el oyente puede considerar al “auto negro” (nombrado así por el hablante previo) como el referente de esta expresión; de cualquier modo, este nuevo hablante no es consciente de la existencia de un pacto previo y por lo tanto puede nombrar a los referentes como quiera, a diferencia del hablante original que está restringido por el pacto conceptual (Metzing y Brennan, 2003). En resumen, la expectativa de consistencia se sostendría en el *common ground* vías pactos conceptuales, por lo que se espera solo del hablante específico con quien se ha generado el pacto.

Sin embargo, la evidencia empírica muestra que la expectativa de consistencia es genérica, pues no depende de la existencia de pactos conceptuales. Se ha mostrado que el oyente interpretará el auto negro una segunda vez de la misma manera, independiente de si ha sido enunciado por el hablante que previamente nombró ese auto así o por un hablante distinto que nunca antes se había referido a ese auto –y que además no estaba al tanto de que para el oyente ese auto ya había sido nombrado–. Pero más contra-intuitivo aún, si el nuevo hablante utiliza una expresión nueva (como “el auto grande”), el oyente evitará mapear este nuevo nombre al auto previamente –pero por otro hablante– denominado “el auto negro”. Esto, aun cuando ambas expresiones son igualmente buenas para el referente en cuestión (un auto de color negro grande). Es decir, el oyente espera consistencia con patrones específicos de referencia de cualquier hablante. Si es una expresión ya conocida, está siendo utilizada para referirse a un objeto ya mencionado con ese nombre. Si es una expresión nueva, el referente es uno no nombrado aún.

La información del hablante genérico, codificada en forma de expectativa pragmática, influye en la interpretación de una referencia a través de mecanismos inferenciales simples que producen interpretaciones suficientemente buenas y rápidas. Son buenas y rápidas porque descansan en información sedimentada y permanente, contenida en una estructura de conocimiento estable. A diferencia de las interpretaciones egocéntricas, las interpretaciones que se derivan de estas heurísticas refieren a un otro, pero genérico, es decir, del cual no es necesario inferir sus estados epistémicos. Estas interpretaciones genéricas son las más veces suficientes para lograr coordinar acciones y lograr, en apariencia, comunicarse.

Hacia una síntesis

He descrito el diseño de los mecanismos que explicarían la interpretación de la referencia definida en contextos conversacionales, tanto en las estructuras de información como en los procesos que transforman esta información. La primera estructura descrita es una forma particular de conocimiento mutuo, denominada *common ground*, que tiene una naturaleza contingente pues codifica el conocimiento compartido y que se sabe compartido entre los interlocutores. Es decir, su información puede ser descrita como representaciones de estados mentales de un hablante específico. Es la naturaleza contingente lo que hace que sea esta estructura lo que hace a cada conversación una instancia comunicativa única. La información de esta estructura es procesada por un mecanismo de monitoreo de la comprensión, que de manera exhaustiva va corrigiendo interpretaciones preliminares si se detecta alguna inconsistencia con el estado de cosas que pueda provocar un malentendido. Un ejemplo de un estado de cosas imposible es un hablante refiriendo a objetos que no sabe que existen. La segunda estructura codifica patrones generales de conducta comunicativa que devienen en expectativas sobre cómo los interlocutores deben comportarse en una interacción conversacional. Estas expectativas se aplican a un interlocutor no porque sea uno en específico, sino porque es un interlocutor que pertenece a la categoría de agentes de los cuales cabe esperar esta conducta: un hablante genérico. En tanto codifica lo que hay de común en todas las conversaciones, esta estructura de conocimiento es lo que hace de una conversación todas las conversaciones anteriores. Estas expectativas alimentan mecanismos simples de inferencia que producen interpretaciones rápidas y muchas veces precisas, pero que no consideran los estados epistémicos del hablante. Son así interpretaciones rápidas que pueden ser corregidas por los mecanismos de monitoreo de la comprensión.

Palabras finales

La comunicación durante una conversación está cruzada por incertidumbre. Los oyentes no pueden nunca estar seguros que lo que están considerando como lo que el hablante quiere decir es efectivamente lo que intenta

comunicar. Y esto no es solo el caso para actos conversacionales complejos y elaborados, como la ironía, la metáfora o los actos de habla indirectos. Esta incertidumbre está presente también –y no en menor medida– en la referencia, aun cuando esta pareciera a simple vista uno de los actos conversacionales más simples. Así, reconocer cuál es la entidad que un hablante intenta traer a la atención conjunta es una acción rodeada de incertidumbre: un mismo objeto puede ser nombrado utilizando distintas formas lingüísticas y una misma forma lingüística puede ser utilizada para referirse a distintos objetos. Esto es cierto incluso si en vez de una forma lingüística se utiliza un gesto como el apuntar. Si alguien apunta hacia donde se encuentra un perro, no es del todo claro que está refiriéndose a las patas del perro, al color del perro, a la forma de caminar del perro o a todas las anteriores (Quine, 1960). A pesar de esta incertidumbre, las más de las veces logramos reconocer la entidad referida, y lo hacemos sin mayor esfuerzo.

A lo largo de este escrito he intentado exponer cómo el fenómeno de la interpretación de una referencia durante una conversación puede ser abordado científicamente, incluso si su primer tratamiento fue filosófico. La particularidad de una abordaje científico es que tiene por objetivo describir el cómo funciona un sistema, a diferencia de un abordaje filosófico, que tiene por objetivo determinar las condiciones bajo las cuales el fenómeno puede realizarse. Así, dar una explicación científica pasa necesariamente por describir los mecanismos causales que explican su funcionamiento, que es otro modo de decir que requiere describir el diseño de un sistema.

Pero como toda apuesta científica, existen muchas otras posibles maneras de explicar este fenómeno. En otras palabras, se puede describir otro diseño que produzca el mismo efecto. He restringido este ensayo a partir de la evidencia existente y solo desde una aproximación particular. El lector puede juzgar si lo expuesto tiene coherencia y así sacar sus propias conclusiones. Lo que sí espero haya sido trasmitido, es que la estrategia de diseño para explicar la interpretación de la referencia en contextos conversacionales es un camino interesante a seguir para quien quiere dar una explicación naturalista de fenómenos mentales.

Agradecimientos

Parte de este trabajo fue financiado por el proyecto Fondecyt Iniciación N°11100226.

Referencias

Bechtel, W. (2008). *Mental mechanisms: philosophical perspectives on cognitive neuroscience*. London: Taylor y Francis.

Brennan, S. E. y Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, y Cognition* 22: 1482-1493.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H. y Marshall, C. R. (1981). Definite reference and mutual knowledge. En: A. K. Joshe, B. L. Webber y I. A. Sag (Eds.). *Elements of discourse understanding*. Cambridge: Cambridge University Press, pp. 10-61.

Clark, H. H., Schreuder, R. y Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior* 22: 245-258.

Clark, H. H. y Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22: 1-39.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy* 68: 87-106.

Garrod, S. y Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27(2): 181-218.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gigerenzer, G., Todd, P. M. y ABC Research Group (1999). *Simple heuristics that makes us smart*. Oxford: Oxford University Press.

Grice, H. P. (1957). Meaning. *The Philosophical Review* 66: 377-388.

—— (1975). Logic and conversation. En: S. Davis (Ed.). *Pragmatics: A reader*. New York: Oxford University Press, pp. 305-315.

Keysar, B; Lin, S. y Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89(1): 25-41.

Keysar, B; Barr, D. J; Balin, J. A. y Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Science* 11: 32-38.

Keysar, B; Barr, D. J; Balin, J. A. y Paek, T. S. (1998). Definite reference and mutual knowledge: process models of common ground in comprehension. *Journal of Memory and Language* 39: 1-20.

Kronmüller, E; Morisseau, T. y Noveck, I. A. (2014). Show me the pragmatic contribution: a developmental investigation of contrastive inference. *Journal of child language* 41(05): 985-1014.

Kronmüller, E. y Barr, D. J. (2007). Perspective-free pragmatics: broken precedents and the recovery-

from-preemption hypothesis. *Journal of Memory and Language* 56: 436-455.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT press.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Metzing, C. y Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language* 49: 201-213.

Quine, W. (1960). *Word and object.* Cambridge: MIT Press.

Schiffer, S. R. (1972). *Meaning*. London: Oxford University Press.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of psycholinguistic research* 32(1): 3-23.

Senft, G. (2014). *Understanding language use: pragmatics*. London: Routledge.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review* 63(2): 129-138.

Sperber, D. y Wilson, D. (1986). *Relevance: communication and cognition*. Oxford: Blackwell.

Strawson, P. F. (1950). On referring. *Mind* 59: 320-344.

CUANDO LA MENTE SE ALTERA

CAPÍTULO IX

La aproximación fenomenológica en psiquiatría

Otto Dörr Zegers

Introducción

Es prácticamente imposible resumir brevemente los principios y métodos de la fenomenología y su enorme influencia sobre las ciencias humanas, así como los rasgos característicos de una de sus derivaciones: la psiquiatría fenomenológica. Solo podemos destacar los elementos que parecen ser fundamentales para nuestra comprensión de su rol en la historia de la psiquiatría. Así, se considera que la fenomenología y el movimiento fenomenológico fueron fundados por el filósofo alemán Edmund Husserl a principios del siglo XX. Husserl estudió primero matemáticas y luego filosofía con Franz Brentano, de quien adoptó el concepto de la intencionalidad de la conciencia. Para Brentano (1925), podemos definir los fenómenos psíquicos diciendo que son aquellos fenómenos que, precisamente como intención, contienen un objeto en ellos mismos. Luego Husserl expresa lo mismo en sus propias palabras: “En los simples actos de percibir estamos dirigidos a las cosas percibidas, en el recordar, a las recordadas, en el pensar, a los pensamientos, en el evaluar, a los valores, en el desear, a los objetivos y perspectivas” (1962, p. 237). La realidad está dada en primer lugar como “noema” de una “noesis”, como el contenido de nuestros actos conscientes.

La filosofía tiene que desarrollar una aproximación directa e inmediata a los objetos y evitar los muchos prejuicios que han caracterizado a la historia y al conocimiento humano. El mandato de Husserl “a las cosas mismas” no significa simplemente rechazar todos los dogmas tradicionales, es más bien una invitación a sumergirse totalmente en las cosas. Eugen Fink, seguidor de Husserl, escribió: “[es un asunto de] una inmersión en los auténticos

problemas como tales… (sin) luchar con pequeñas evidencias" (1958, p. 3). Un regreso a las cosas como ellas son en sí mismas no significa que uno tome los hechos concretos a valor nominal, como hace, por ejemplo, el realismo. Al contrario, uno debería llegar más allá de aquello que se muestra a sí mismo. Ahora bien, la fenomenología no se limita a investigar los datos de la experiencia sensorial, sino que acepta igualmente datos no sensoriales (como los valores y las relaciones) que son captados por la intuición en forma directa.

La experiencia fenomenológica no es por cierto una experiencia natural, aunque procede de esta. En la vida diaria uno se encuentra en una actitud natural ingenuamente orientada hacia el mundo de los objetos. Esto no nos lleva, sin embargo, ni al conocimiento ni menos aún al conocimiento científico. Para que esta actitud natural nos conduzca al conocimiento científico, necesitamos reducir el objeto a uno solo de sus aspectos. Así, por ejemplo, cuando el químico se aproxima al agua, él reduce todos sus significados a la mera composición molecular: dos átomos de hidrógeno y uno de oxígeno. En esencia, el científico natural proyecta la teoría físico-química de la realidad sobre la totalidad del fenómeno, descartando todos los otros elementos constitutivos del objeto real. En su aproximación al objeto el químico no considera la capacidad del agua para saciar la sed o para hacer que los campos produzcan frutos y nunca tomará en cuenta los simbolismos de la profundidad del mar, la importancia de las nubes o la belleza de un lago. En contraste, cuando el fenomenólogo adopta una actitud reflexiva dirige su atención a la totalidad de las muchas formas en que un objeto es percibido en la conciencia. Broekman, un destacado alumno de Husserl, afirma en este contexto: "Esta división fundamental en dos actitudes (la natural y la fenomenológica) y sus respectivos campos es un elemento principal que puede ser encontrado en toda fenomenología" (1965, p. 165). La actitud fenomenológica intenta, como lo formula Blankenburg (1962):

> Ser lo más abierta posible a las diferentes formas de ser del objeto, es decir, intenta ser (aún) "más natural" que la propia experiencia natural. Pero por otra parte, también intenta ser "más científica" que la experiencia científica, en la medida en que no se limita a un aspecto particular, sino que transforma en su objeto (la totalidad de) las formas de ser de lo que nos enfrenta (p. 413).

En otras palabras: en toda experiencia real experimentamos más de lo

que la percepción del mero objeto nos proporciona. Esto fue formulado en forma brillante cien años antes de Husserl por Goethe en uno de sus aforismos: “La experiencia es siempre solo la mitad de la experiencia” (1966, p. 687). Siempre vivimos más de lo que vivimos, experimentamos más de lo que experimentamos y explorar esta otra parte es la gran tarea de la fenomenología. El principio de Goethe, en sí mismo tan similar a las postulaciones de Husserl, nos lleva directamente a la obra de Marcel Proust. El profundo significado de su novela *En busca del tiempo perdido* está en el rescate de todo lo que él experimentó en el pasado y vivió en ese momento sin casi tomar conciencia de ello. Las principales características de su obra son similares a las etapas fundamentales del método fenomenológico: una apertura total a la realidad, una actitud reflexiva que percibe la realidad como entregada a la conciencia y una eliminación progresiva de todos los presupuestos, prejuicios y elementos accidentales como un instrumento para lograr una visión de la esencia de lo que se experimenta.

El primero que aplicó el método fenomenológico a la psiquiatría fue el médico y filósofo alemán Karl Jaspers (1963a, 1963b). En la perspectiva de Jaspers la fenomenología es una “psicología descriptiva”. Esta está firmemente relacionada con hechos y entrega una descripción imparcial de las experiencias de los pacientes. Jaspers combinó un llamado a la empatía (*Einfühlen*) junto con la inmediata comprensión (*Verstehen*) de la otra persona, sin explotar técnicas propiamente fenomenológicas como la reducción, la variación libre o la intuición de esencias. Sin embargo, nosotros no compartimos la
opinión de Berríos (1989), quien destacó demasiado el contraste entre Husserl y Jaspers. Las diferencias entre ambos no son tan extremas, al menos en lo que se refiere a la aplicación de la fenomenología a la psiquiatría. Algunas afirmaciones de Jaspers pueden ilustrar lo antedicho: “La fenomenología se relaciona con lo que es experimentado como real; observa la mente ‘desde dentro’ a través de la inmediata representación (*Vergegenwärtigung*)” (1963a, p. 326). Así, los puntos de partida de Jaspers fueron en realidad similares a los de Husserl: el regreso a las cosas, la intencionalidad de los fenómenos psíquicos y el cambio desde una actitud natural a la fenomenológica. Autores como Osborne Wiggins y Michael Schwartz (1992, 1997a) han expresado recientemente una opinión análoga al respecto.

El segundo psiquiatra que aplicó el método fenomenológico a la psiquiatría es el suizo-alemán Ludwig Binswanger. En su artículo “Sobre la fenomenología” ya sugiere que “la intuición de las esencias es el proceso fundamental del método fenomenológico (en psiquiatría)” (1947, p. 19). Luego la influencia de *Ser y tiempo* de Heidegger se hizo irresistible y causó su conversión al existencialismo. A través de él la forma de ser y la biografía de los pacientes psiquiátricos se colocaron dentro del dominio de la descripción que hace Heidegger de la existencia humana como “ser-en-el-mundo”, de modo que “No decimos que ‘las enfermedades mentales sean enfermedades del cerebro’, sino que en las enfermedades mentales encontramos modificaciones de la estructura esencial y elementos estructurales del ser-en-el-mundo como trascendencia” (1947, p. 194). El análisis existencial explora por lo tanto en qué formas un *Dasein* “descubre”, “proyecta” o “se abre” hacia un mundo (1957). Ahora intentaremos resumir sus hallazgos pertenecientes a la historia vital y el mundo de los pacientes esquizofrénicos:

1. En primer lugar, el desarrollo de una existencia esquizofrénica se caracteriza por “la ruptura de la coherencia de la experiencia natural”. Binswanger persigue este patrón de “incoherencia” hasta la misma infancia de sus pacientes. Por esta razón sus vidas son a menudo dolorosas y difíciles, pues ellos están permanentemente buscando la forma cómo restablecer ese orden perdido en forma tan temprana.
2. Otro elemento esencial de la existencia esquizofrénica es la escisión en alternativas irreconciliables. La alternativa dominante es generalmente un ideal excéntrico. Esta tendencia a la escisión aparece a menudo en el período previo a la enfermedad en los diferentes ámbitos de la vida, incluida la relación con el propio cuerpo.
3. El tercer rasgo característico es el ocultamiento del lado rechazado de la alternativa a favor de la absoluta supremacía del ideal excéntrico. En esta perspectiva uno podría entender la relación entre esquizofrenia y manierismo, porque el ocultamiento de la alternativa temida se logra por medio de modos de expresión arbitrarios, desfigurados e innecesariamente detallados, v. gr., en el lenguaje, en el dibujo, etc.
4. La última característica es su tendencia al “aniquilamiento”, que se hace manifiesta como la culminación de tensiones conflictivas y la retirada del

mundo. Este es el ansiado fin, que puede consumarse ya sea en la forma de un autismo pasivo, un delirio crónico o el suicidio.

En sus últimos dos libros, *Melancolía y manía* (1960) y *Delirio* (1965), Binswanger abandonó estos análisis biográficos tan ligados a la obra de Heidegger para volver a la fenomenología trascendental de Husserl. Intentó describir los modos de constitución de los mundos patológicos como perturbaciones de la conciencia trascendental. Al comprender el tiempo desde la intencionalidad Husserl había diferenciado entre *protentio* (futuro), *retentio* (pasado) y *praesentatio* (presente). En el caso de la depresión endógena Binswanger encontró una profunda alteración en la constitución de estos éxtasis: La *protentio* está invadida por elementos del pasado (en un delirio depresivo una mera posibilidad, que yo me arruine, por ejemplo, se transforma en algo ya ocurrido), mientras la *retentio* es invadida por elementos del futuro (como ocurre en los autorreproches, donde lo que ya es definitivamente pasado es vivido todavía como si fuera una posibilidad). En la manía Binswanger encontró un aflojamiento de los tres momentos constitutivos de la temporalidad. El futuro flota en el aire sin ningún anclaje en el presente ni tampoco en el pasado y este deja de ser necesario para sostener el presente o el futuro. Por lo tanto, un maníaco no necesita haber estudiado medicina para considerarse un gran médico. Binswanger también encontró una falla en la constitución de la interpersonalidad, tema complejo sobre el cual no nos podemos extender en este contexto (Szilasi, 1959).

Esta vuelta de Binswanger a la fenomenología pura y trascendental ha sido criticada por algunos de sus seguidores, como K. P. Kisker (1961) y H. Tellenbach (1962). Ambos dudan de que este regreso a la conciencia trascendental y sus distintas constituciones signifique un verdadero progreso para la fundamentación científica de la psiquiatría. Ellos estiman como una pérdida el que Binswanger haya abandonado el análisis existencial, con su acento puesto en la historicidad, para dar paso al estudio de estas estructuras tan a-históricas, como son las de la conciencia trascendental. Sin embargo, no se puede negar que la aplicación de la fenomenología trascendental de Husserl también puede rendir frutos a la psiquiatría. Blankenburg ha demostrado esto ampliamente (1962, 1971) y recientes esfuerzos por construir un vínculo entre la fenomenología y la psicopatología a la luz de tendencias emergentes en la neurobiología, así

como en la filosofía de la conciencia (Mishara y Shwartz, 1997) apuntan en la misma dirección.

Pero otros autores europeos de la época también recibieron la influencia de Husserl y de Heidegger. Es el caso de V. E. von Gebsattel, E. Straus y J. Zutt en Alemania, E. Minkowski en Francia y J. J. López-Ibor en España. Von Gebsattel realizó estudios sobre el mundo fóbico y el mundo obsesivo, considerados hoy imprescindibles. Su descripción de la hipersensibilidad del fóbico a la dimensión de la amplitud del espacio (1954, p. 63), así como su elaboración del tema de la descomposición, la suciedad y la impureza en tanto elementos constitutivos y determinantes del mundo obsesivo (1954, p. 122), han de ser consideradas fundamentales. Straus, por su parte, hizo importantes aportes al tema de la espacialidad humana en general (1960, p. 141) y al problema de la temporalidad en la depresión (1960:126). Notable es también la obra de Minkowski (1933), *Le temps vécu*, que no solo representa un estudio muy profundo de la temporalidad de las psicosis, sino también un aporte definitivo a la comprensión del fenómeno del "autismo". La obra de López-Ibor, de gran repercusión en el mundo hispano-parlante, es un buen ejemplo de cómo la intuición fenomenológica va por delante de los resultados de la investigación empírica. Ya en 1950 López-Ibor (*La angustia vital*) definió el carácter "endógeno" de los cuadros entonces llamados "neuróticos". La desaparición del concepto de neurosis desde el DSM III (1980) y la ampliación del concepto de depresión en las últimas décadas han confirmado sus intuiciones fenomenológicas.

Otro autor que ha hecho una gran contribución a la orientación fenomenológica en psiquiatría es Jürg Zutt. Nos detendremos brevemente en su obra, por cuanto su análisis del fenómeno alucinatorio (1963) es uno de los ejemplos más claros de cómo una aproximación fenomenológica puede contribuir al conocimiento de la esencia de un fenómeno psicopatológico. Muchos enfermos escuchan voces, las que tradicionalmente han sido interpretadas como "percepciones engañosas del oído". Si el órgano sensorial "engañado" es el oído, se piensa que la función alterada es la vía acústica y de hecho se realizaron en la primera mitad del siglo XX muchas investigaciones empíricas orientadas a descubrir esta alteración, obviamente sin éxito. Por otra parte, ya a los autores clásicos les llamaba la atención el que los esquizofrénicos escuchen voces, pero no vean visiones. ¿Por qué esta discrepancia? Esta aproximación a la alucinación y

su discrepancia es coherente con la teoría del ser humano como organismo, en el sentido de la anatomía y la fisiología. Pero existen otras teorías, como la psicoanalítica, en la cual la alucinación no es vista como perturbación auditiva, sino como producto del repliegue de un Yo débil frente a contenidos angustiosos de la conciencia. En ambos casos, tanto en la teoría organicista como en la psicoanalítica, el síntoma "oír voces" no es estudiado en sí mismo, sino más bien "encubierto", al no dejarlo hablar desde y por sí mismo. Este autor ha contribuido al análisis de Zutt, al comparar la alucinación esquizofrénica con la del paciente alcohólico (1963, p. 273). Este escucha hablar a personas inexistentes que de pronto se presentan en el espacio real ("ahí", "detrás de la puerta", "en la calle", etc.) y frente a las cuales él reacciona, mientras que el paciente esquizofrénico al escucharlas se siente paralogizado y avasallado y no actúa de acuerdo a lo escuchado. La alucinación del alcohólico es un activo percibir, porque su percepción implica un identificar ("es la policía"), un localizar ("ya atravesaron la puerta") y un actuar (el paciente se esconde debajo de la cama). En cambio la alucinación del esquizofrénico es pura pasividad e indefensión, por cuanto él, propiamente, no escucha voces (como el alcohólico), sino que es hablado. Y no bien hemos llegado a esta fórmula, caemos en la cuenta que la mencionada discrepancia ha desaparecido, porque el esquizofrénico, así como es hablado, es mirado y es tocado; vale decir, la correspondencia no es voz-visión, sino voz-mirada. La perturbación fundamental en esta enfermedad no radica entonces en la percepción, sino en ese fenómeno mucho más complejo y abarcador que es la interpersonalidad.

El nombramiento de W. von Baeyer y posteriormente de W. Janzarik como catedrático de Psiquiatría de Heidelberg inició una nueva era en la psiquiatría fenomenológica, asociada, además de los recién mencionados, a los nombres de W. Blankenburg, H. Haefner, K. P. Kisker y H. Tellenbach. En esa época muchos psiquiatras de Alemania y del extranjero recibieron su entrenamiento en esta clínica universitaria o recibieron su influencia. La mayoría de ellos permanecen vinculados a la fenomenología. Ese es el caso, entre otros, de A. Kraus, H. Lang, Ch. Mundt, U. W. Peters, M. Schmidt-Degenhard y M. Spitzer en Alemania, de Bin Kimura en Japón, de R. Ballbé en Argentina, de G. Figueroa, C. Ojeda, R. Parada y este autor en Chile. Como no es posible referirse a la obra de cada uno de ellos, nos

detendremos solo en Hubertus Tellenbach y Wolfgang Blankenburg, cuyas obras tuvieron la mayor repercusión fuera de Alemania.

La obra de Tellenbach es prodigiosa, porque no solo nos muestra la fecundidad del pensamiento fenomenológico en la comprensión de la esencia de las enfermedades psíquicas, sino la clara posibilidad de su aplicación a la práctica clínica. El campo que más investigó fue el de la enfermedad depresiva (1983), pero también hizo aportes importantes a la comprensión del delirio, de "lo atmosférico" y de la figura del padre. También realizó análisis hermenéuticos de personajes literarios y religiosos, como el Rey Edipo, Hamlet, Otelo, el Rey Lear y el Príncipe Myshkin (1992), en la búsqueda de lo que sus autores nos pueden enseñar sobre los fenómenos psicopatológicos como el delirio, la melancolía, los celos, la demencia y la epilepsia. Tellenbach también describió un tipo humano específico, el "*typus melancholicus*", que sería proclive a la depresión unipolar. Este tipo estaría caracterizado fundamentalmente por un rasgo que él denomina con el neologismo "ordenalidad", que significa un quedar fijado de tal modo a los órdenes naturales de la vida (el habitar, el matrimonio, la maternidad o paternidad, el trabajo, etc.), que los riesgos que esta implica y/o la imposibilidad de mantener ese orden en la forma acostumbrada pueden transformarse en una amenaza existencial y una puesta en marcha del cambio endógeno hacia la depresión ("*endokinesis*"). Quizás si lo más interesante de su concepción sea que ella nos permite descubrir la perfecta coherencia entre personalidad pre-depresiva, situación desencadenante y sintomatología. Su visión abarcadora y visionaria le permitió afirmar ya en 1961 y cuando se distinguían alrededor de veinte tipos de depresiones diferentes, la existencia de un solo sindrome monopolar "endoreactivo". Con ello Tellenbach se
adelantó casi veinte años a la descripción de la "depresión mayor" del DSM III. Posteriormente von Zerssen *et al.*, (1982) en Alemania, junto a Anneliese Dörr y Sandra Viani (1999) en Chile demostraron empíricamente la validez de las intuiciones de Tellenbach.

Blankenburg, por su parte, intentó primero aplicar la fenomenología trascendental de Husserl a un caso de un paciente con humor delirante (*Wahnstimmung*). Un joven estudiante ante la vitrina de una librería tuvo la sorprendente sensación de que todos los libros estaban expuestos ahí para él: "Era como un reproche. Quizás eso significaba que yo tenía que ser más

estudioso en la universidad" (1962, p. 413). El punto de partida de Blankenburg es justamente su propia experiencia frente al enfermo delirante y así descubre que su capacidad de asombrarse ha sido "desbordada", sobrepasada. Intenta entonces penetrar más profundamente en lo que significa "asombrarse" y "desbordarse". En el asombrarse "lo que emana del mundo (*Zuwurf*) alcanza todavía a ser contenido en un proyecto (*Entwurf*) más amplio de la existencia humana", lo que no ocurre en el "desbordarse" (1962, p. 415). Y este es el caso del joven estudiante: de los libros extraídos de su neutralidad se ha desprendido algo que acosa al paciente, que lo deja al descubierto y sin ninguna protección. En consecuencia, se pierde la coherencia o lógica de su experiencia natural, lo que vendría a ser el germen de lo que en casos más grave conocemos como influencia del pensamiento (*Gedankenbeeinflussung*).

Una de las más importantes contribuciones de Blankenburg es su concepto de "la pérdida de la evidencia natural" (*der Verlust der natürlichen Selbstverständlichkeit*) en la esquizofrenia. Él observó este fenómeno en una serie de pacientes con un elevado nivel intelectual y gran capacidad de auto-reflexión. Una de sus pacientes decía: "me falta algo, algo pequeño, pero tan importante que sin ello no se puede vivir" (1971, p. 60). Otro pensaba que la salud y la felicidad solo eran posibles porque había un "algo" del que habitualmente no se está consciente. El mismo paciente terminaba diciendo que "este misterioso 'algo' parece resistirse tenazmente a la conciencia" (1971, p. 61). Las personas sanas no reflexionan sobre el *common sense* y la comprensión de la vida diaria, pero los esquizofrénicos se preocupan de ello porque lo han perdido. Sin embargo, hay una situación –la *epojé* fenomenológica– en la cual las personas sanas están muy cerca de lo que les ocurre a los esquizofrénicos. La *epojé* consiste justamente "en un desprenderse de las evidencias de la vida cotidiana, de ese vivir, actuar y opinar simple e ingenuamente, en el que estamos enraizados en el mundo de la vida" (1971, p. 64), para adoptar una actitud reflexiva y distante. La resistencia que el fenomenólogo experimenta al adoptar esta actitud es aquello de lo cual carece el paciente esquizofrénico. Se podría decir que este vive, en contraste con el fenomenólogo, en una *epojé* descontrolada y permanente.

En forma paralela al desarrollo de la psiquiatría fenomenológica en Alemania surgió también un movimiento similar en Francia,

fundamentalmente en torno a dos polos, uno en París, encabezado por Yves Pelicier y otro en Marsella, vinculado a los nombres de Sutter, Tatossian, Azorin, Giudicelli y Naudin. Francia constituye hoy uno de los más importantes centros de investigación fenomenológica, una de cuyas expresiones es la revista *L'art du comprendre*, que dirigen Philippe Forget y Georges Charbonneau. La obra de Pelicier es muy extensa y se caracteriza por su punto de partida clínico y su importante preocupación ética. Casi no hay tema de la psiquiatría que le fuera ajeno y sobre cada uno de ellos (angustia, depresión, adicciones, demencia, enfermedades psicosomáticas) hizo aportes significativos (Pelicier, 1987). Se podría decir que su método se ubica en una posición intermedia entre el de Jaspers y el de Husserl. Él se atiene, como Jaspers, rigurosamente a la experiencia del paciente concreto, pero como Husserl logra determinar características esenciales, elementos invariantes dentro de la complejidad de los hechos psicopatológicos, aunque sin pretender preguntarse por la constitución de los actos de conciencia. La obra de Tatossian (1979) *Phénoménologie des psychoses* significó la primera sistematización de los conocimientos en el campo de las psicosis y, al mismo tiempo, la introducción en Francia de todos los autores alemanes desde Binswanger hasta Blankenburg. No podemos terminar esta breve alusión a la fenomenología francesa sin mencionar una obra realmente extraordinaria, cual es *Phénoménologie et psychiatrie: le voix et la chose,* de Jean Naudin (1997) y que representa el más acabado análisis fenomenológico del fenómeno alucinatorio. Para él, igual que para Zutt, la alucinación no es un trastorno de la percepción, sino de la intencionalidad misma: "(La alucinación) representa una deshumanización sin límite que pone al desnudo la fenomenalidad misma del proceso intencional" (p. 323). Y más adelante agrega: "Transparente y unívoco, el Otro alucinado se puede definir por su falta de reciprocidad. No es un *alter ego*, puesto que él es vivido en la inmediatez… Todo ocurre como si se hubiera invertido ese movimiento perspectivista que, en la percepción, nos entrega la cosa siempre en esquemas sucesivos: la idealidad está aquí dada de un golpe" (1997, p. 323).

Pero no solo en Francia la fenomenología psiquiátrica se encuentra en pleno desarrollo. La revista *Philosophy, Psychiatry and Psychology* ha reunido importantes autores de Inglaterra y los Estados Unidos, entre los que cabe destacar a K. W. M. Fulford (UK) y a M. Schwartz (EE.UU.).

Tampoco habría que olvidar que la importante revista *Current Opinion in Psychiatry*, editada por la WPA y que representa un verdadero foro donde se discuten los temas psiquiátricos al más alto nivel, está publicando cada día con mayor frecuencia contribuciones fenomenológicas.

Con respecto al futuro, nosotros compartimos las ideas planteadas por O. Wiggins y M. Schwartz (1997b). La fenomenología puede contribuir a la psiquiatría, en primer lugar, en el desarrollo de una antropología que permita comprender las enfermedades mentales más allá del dualismo psico-físico; y en segundo lugar y por tratarse de un método científico riguroso, puede desarrollar conceptos psicopatológicos que permitan, por ejemplo, diferenciar el delirio en una esquizofrenia, en una manía o en una psicosis exógena.

Para terminar quisiéramos insinuar una crítica a la fenomenología psiquiátrica. Esta se refiere al hecho que la aplicación del método fenomenológico en psiquiatría lleva siempre a la constatación de una "pérdida" de algo que el hombre sano posee –problema también señalado por Broekman (1998)– como por ejemplo, "la pérdida de la evidencia natural" (Blankenburg) o "la pérdida de la consecuencia de la experiencia natural" (Binswanger). El buscar la mera deficiencia puede hacer también que sus determinaciones parezcan ser solo traducciones de conceptos psicopatológicos en otro lenguaje, como cuando en lugar de "devastamiento" esquizofrénico (*Verblödung*) se habla de "retraimiento existencial" (*Rückzug*). Nosotros vemos un solo camino para superar este escollo y ese es la profundización dialéctica de la actitud fenomenológica hasta el punto de poder comprender las desviaciones psicopatológicas positivamente. Esto quiere decir que ellas deben ser vistas no como meras deficiencias, "sino como la emancipación de momentos particulares integrantes –y en el sano, al mismo tiempo, integrados– de la realidad humana" (Blankenburg, 1985). En el caso del delirio (fenómeno que pertenece al ámbito de los actos intencionales que fijan la realidad y se concretan, en último término, en juicios), Blankenburg ha demostrado cómo la normal constitución de un juicio sobre la realidad se muestra ella misma como una formación delirante *in statu nascendi*, la que está siendo permanentemente corregida o abolida. En el delirio se produce una emancipación y radicalización de posibilidades tan humanas como la capacidad de ascender de lo particular a lo general, de imaginar

posibilidades que están más allá de lo real o de buscar la verdad sin transiciones. Nosotros hemos intentado mostrar cuán estrecho es el vínculo entre delirio y verdad (Dörr-Zegers 1997, p. 128). En rigor, el delirio tiene casi todas las atribuciones de una verdad que ilumina (en el sentido de la *aletheia* griega), solo que la radicalidad con que la vive el enfermo delirante termina encegueciéndolo y aislándolo.

Pero esta perspectiva fenomenológico-dialéctica no solo tiene un interés psicopatológico en el sentido de permitir comprender esos modos de vivenciar y comportarse desde sí mismos y no como meras desviaciones con respecto a una supuesta norma. Está también la perspectiva de la positividad de lo negativo desarrollada originalmente por Hegel y aplicada a la psiquiatría primero por Blankenburg (1981) y luego por el suscrito (Dörr-Zegers, 1990). Así, podría dirigirse la mirada no solo a las deficiencias del esquizofrénico (lo negativo), sino a muchos aspectos positivos que le son propios, como su vuelo metafísico, su capacidad de percibir matices de la realidad que son ajenos al hombre corriente, o su búsqueda de la verdad sin concesiones (Dörr-Zegers y Tellenbach, 1980, p. 128). El mismo enfoque podría adoptarse en el caso de un paciente depresivo, histérico u obsesivo-compulsivo. La personalidad depresiva muestra lealtad a las normas sociales, alto rendimiento en el trabajo y olvido de sí mismo. La personalidad histérica se adapta a cualquier contexto, posee una gran capacidad para apasionarse y es un destacado animador. El paciente obsesivo muestra una fuerte capacidad para realizar proyectos, una sobriedad en la apariencia y un trato cuidadoso con las otras personas.

Ahora bien, el interés de este planteamiento no es solo teórico, sino también práctico. Al no tomar lo negativo simplemente como tal, el psiquiatra se verá obligado a ampliar su horizonte de comprensión, a cambiar marcos de referencia y a cuestionar el concepto tradicional de enfermedad. Además, a esta mirada se nos presentan los distintos cuadros patológicos ordenados en polaridades (esquizofrenia-depresión; histeria-obsesión), con múltiples transiciones entre ellas, lo que corresponde mucho más a la experiencia clínica que el empleo de categorías nosológicas absolutas. El ver a una estructura psicopatológica como el lado positivo de la otra nos abre un camino de acción terapéutica privilegiado, cual es el evitar una mera adaptación a ese inexistente término medio, procurando

hacer tomar conciencia al paciente de la positividad de sus rasgos pretendidamente anormales, pero de tal modo que empiece un camino en la dirección contraria, hacia su polo opuesto, v. gr. desde el polo esquizofrénico al depresivo y viceversa. Al hacerlo, la actitud fenomenológica ayuda al paciente a acercarse a la medida griega, pues, como dice la vieja sabiduría de Heráclito en el Fragmento
Nº 8: “Lo que se opone se une; de las cosas diferentes nace la más bella armonía” (1967, p. 8).

Referencias

Berríos, G. E. (1989). What is phenomenology? *Journal of the Royal Society of Medicine* 82, pp. 425-428.

Binswanger, L. (1947). *Ausgewählte Vorträge und Aufsätze I*. Francke: Bern.

—— (1957). *Schizophrenie*. Pfullingen: Neske.

—— (1960). *Melancholie und Manie*. Pfullingen: Neske.

——(1965). *Wahn*. Pfullingen: Neske.

Blankenburg, W. (1962). Aus dem phänomenologischen. Erfahrungsfeld innerhalb der Psychiatrie. Schweizer Archiv für Neurologie. *Neurochirurgie und Psychiatrie* 90, pp. 412-421.

—— (1971). *Der Verlust der natürlichen Selbstverständlichkeit*. Stuttgart: Enke. Existe traducción por Otto Dörr y Elvira Edwards (2013) *La pérdida de la evidencia natural*. Santiago: Ediciones Universidad Diego Portales.

—— (1981). Wie weit reicht die dialektische Betrachtungsweise in der Psychiatrie? *Zeitschrift für Psychiatrie und Psychotherapie* 29, pp. 45-66.

—— (1985). Perspectiva antropológica y analítico-existencial del delirio. *Revista Chilena de Neuropsiquiatría* 23, pp. 165-178.

Broekman, J. M. (1965). Phänomenologisches Denken in Philosophie und Psychiatrie. *Confinia psychiatrica* 8, pp. 165-187.

—— (1998). Schizophrenie als Erkenntnisproblem. *Fundamenta Psychiatrica* 12, pp. 47-52.

Brentano, F. (1925). *Psychologie vom empirischen Standpunkt*. Leipzig: Meiner.

Dörr, A. y Viani, B. (1999). Personalidad premórbida en los distintos cuadros afectivos. Estudio comparativo ex post-facto. *Acta Psiquiat. Psicol. Am. lat.* 45(1), pp. 41.50.

Dörr-Zegers, O. (1997). *Psiquiatría Antropológica* (2nd edición*).* Santiago: Universitaria.

—— (1990). Hacia una concepción dialéctica en psiquiatría. *Actas Luso-Españolas de Neurología y Psiquiatría* 18, pp. 244-257.

Dörr-Zegers, O. y Tellenbach, H. (1980). Differentialphänomenologie des depressiven Syndroms. *Nervenarzt* 51, pp. 113-118.

Fink, E. (1958). *Studien zur Phänomenologie*. Den Haag: Martinus Nijhoff.

Goethe, W. (1966). *Naturwissenschaftliche Schriften II, Sprüche in Prosa (Nº 172)*. Zürich und Stuttgart: Artemis.

Heráclito (1967). *Textos de los grandes filósofos - Edad Antigua*. En: R. Verneaux (Ed.). Barcelona: Editorial Herder.

Husserl, E. (1962). *Husserliana*, Band IX, Phänomenologische Psychologie. Den Haag: Martinus Nijhoff.

Jaspers, K. (1963a). Die phänomenologische Forschungsrichtung in der Psychopathologie. In *Gesammelte Schriften zur Psychopathologie*, pp. 314-328. Springer, Berlin-Göttingen-Heidelberg.

—— (1963b). *Allgemeine Psychopathologie*. Berlin-Göttingen-Heidelberg: Springer.

Kisker, K. P. (1961). Die phänomenologische Wendung Ludwig Binswanger's. *Jahrbuch für Psychologie, Psychotherapie und Medizinische Anthropologie* 8: 143-153.

López-Ibor, J.J. (1950). *La angustia vital*. Madrid: Paz Montalvo.

Minkowski, E. (1933). *Le temps vécu*. Paris: Collection de L"Évolution Psychiatrique.

Mishara, A. y Shwartz, M. A. (1997). Psychopathology in the light of emergent trends in the philosophy of consciousness, neuropsychiatry and phenomenology. *Current Opinion in Psychiatry* 10, pp. 383-389.

Naudin, J. (1997). *Phénoménologie et psychiatrie*. Toulouse: Presses Universitaires du Mirail.

Pelicier, Y. (1987). Asthme, souffle et culture. *Psychologie Médicale* 19, pp. 1849-1850.

Straus, E. (1960). *Psychologie der menschlichen Welt.* Berlin-Göttingen-Heidelberg: Springer.

Szilasi, W. (1959). *Einführung in die Phänomenologie Edmund Husserls*. Tübingen: Niemayer.

Tatossian, A. (1979). *Phénoménologie des psychoses*. Paris: Masson.

Tellenbach, H. (1962). Abbreviatur und Epikritisches zu Ludwig Binswanger's Buch "Melancholie und Manie". *Der Nervenarzt* 33, pp. 515-520.

—— (1983). *Melancholie* (4. Ausgabe). Berlin-Heidelberg-New York-Tokyo: Springer.

—— (1992). *Schwermut, Wahn und Fallsucht in der abendländischen Dichtung*. Hürtgenwald: Guido Pressler.

Von Gebsattel, V. E. (1954). *Prolegomena einer medizinischen Anthropologie*. Berlin-Göttingen-Heidelberg: Springer.

Von Zerssen, D. *et al.*, (1982). Personality and Affective Disorders. In *Handbook of Affective Disorders*, pp. 213-228. New York: Churchill Livingstone.

Wiggins, O.P., Schwartz, M. A., Spitzer, M. (1992). Phenomenological/Descriptive Psychiatry: The Methods of Edmund Husserl and Karl Jaspers. In *Phenomenology, Language and Schizophrenia*

(ed. C. Mundt), pp. 1-25. Springer, New York.

Wiggins, O. P. y Schwartz, M. A. (1997a). Edmund Husserl's Influence on Karl Jaspers" Phenomenology. *Philosophy, Psychiatry, y Psychology* 4, pp. 15-36.

—— (1997b). Psychiatry. In *Encyclopaedia of Phenomenology* (ed. Embree, Lester et al). Boston: Kluwer.

Zutt, J. (1963). *Auf dem Wege zu einer anthropologischen Psychiatrie*. Berlin-Göttingen-Heidelberg: Springer.

CAPÍTULO X

Esquizofrenia y la naturaleza de lo mental: sobre el estudio interdisciplinario de las patologías mentales

Pablo López-Silva

Introducción

La esquizofrenia es uno de los trastornos de la mente humana más debilitantes, y al mismo tiempo, uno de los menos entendidos en la actualidad. Si bien algunas comunidades insisten en definirla como una mera enfermedad con origen biológico a partir de la utilización de estrictos criterios diagnósticos de tipo médico, tal comprensión es actualmente disputada por diversos autores desde distintos puntos de vista (Cooper, 2007; 2013; Pickard, 2009; Poland y Von Eckardt, 2013)[1]. La esquizofrenia afecta a aproximadamente 24 millones de personas a nivel mundial y produce múltiples consecuencias negativas en el bienestar mental y social de los pacientes (Werner, Aviv y Borak, 2008). Por ejemplo, pacientes esquizofrénicos usualmente reportan problemas afectivos, lo que disminuye su funcionalidad social (Marwaha *et al.*, 2013). Junto con esto, no solo los pacientes son afectados por este trastorno, sino que todo su entorno. Así, podemos observar que un entendimiento acabado del fenómeno constituye una contribución para mejorar el bienestar mental y social de las personas que sufren de este trastorno y sus cercanos.

En un nivel conceptual, la esquizofrenia ha sido definida de diversas formas a lo largo de la historia (Bentall, 2003). Por ejemplo, un enfoque socioconstruccionista radical la definiría –implausiblemente– como una mera construcción social dado que las categorías que le dan sentido al supuesto trastorno son siempre fruto de las conversaciones que toman lugar en ciertas redes ideiológicamente definidas[2]. Así, tales categorías solo tendrían sentido dentro de tales comunidades que las legitimizan por medio

de ciertas prácticas sociales (Ver Gergen, 1991). En este enfoque, lo que para algunos sería un “trastorno”, para otros podría ser un “castigo divino” o un acto de “liberación” de ciertas ataduras de carácter político social sin que ninguna de tales categorías posea un valor ontológico objetivo. Sin embargo, dejando a un lado las múltiples críticas a este enfoque (Hacking, 1999; Boghossian, 2006; López-Silva, 2013a), uno los principales errores de esta comprensión de la esquizofrenia es el de focalizar su análisis en aquellos sujetos que diagnostican el fenómeno y no en la comprensión de los esquizofrénicos, quienes a final de cuenta son los que “sufren” de ciertas experiencias conscientes con carácter bizarro y disfuncional (López-Silva, 2013a). Desde este punto de vista, sea cual sea el valor social –percepciones, representaciones, etc. –que una determinada experiencia consciente posee –lo cual claramente es socialmente construido–, la negación de la ocurrencia de una cierta experiencia *qua* experiencia para un sujeto constituye una estrategia implausible. Una cierta comunidad podría entender ciertos fenómenos alucinatorios como “relevaciones divinas”, sin embargo, esto no implica la comprensión y explicación de aquello que es experimentado por el sujeto.

Por otro lado, el paradigma psiquiátrico clásico expresado en la publicación de las diversas versiones de los manuales DSM de la *American Psychiatric Association* y en la publicación de la *International Statistical Classification of Diseases and Related Health Problems* (CIE) ha comúnmente entendido a la esquizofrenia como un “trastorno mental”. Recogiendo algunos aspectos críticos de la psiquiatría fenomenológica de Jaspers y Binswanger, el actual enfoque Danés centrado en la descripción fenomenológica de las experiencias anómalas que caracterizan la esquizofrénica ha definido el fenómeno como un problema al nivel de la consciencia y de la autoconciencia (Parnas y Sass, 2001; Parnas y Handest, 2003; Parnas, Møller, Kircher, Thalbitzer, Jansson, Handest y Zahavi, 2005a).

Sin embargo, es importante notar que tales definiciones –la clásica y la fenomenológica– no solo invitan a la reformulación de ciertas categorías médicas utilizadas en psiquiatría, sino que también invitan a profundos debates a nivel filosófico (ver Fulford, Davies, Gipps, Graham, Sadler, Stanghellini y Thorton, 2013). La comprensión y el diagnóstico de los diversos trastornos mentales existentes nos confronta continuamente con

varias de las preguntas más fundamentales en filosofía de la mente y fenomenología:

· Si el enfoque clásico define la esquizofrenia como un “trastorno mental”, es necesario preguntar: ¿Qué es lo mental? ¿Cómo es que podemos llegar a definir un trastorno como “mental” como diferente a un mero trastorno “físico”? ¿Es acaso lo mental diferente de lo físico? Si lo es, ¿cómo es diferente lo mental de lo físico y cual es la relación entre ambos dominios? Si bien tal distinción parece intuitivamente atractiva, las repuestas a la pregunta por “lo mental” son más que diversas y la discusión sobre la relación entre ambos dominios está lejos de estar resuelta (Armstrong, 1968; Searle, 1980; Jackson, 1986; Crane, 1996). Así, no parece haber consenso sobre qué es lo que muchas veces filósofos, psicólogos y psiquiatras quieren decir por el término “mental” y menos por lo que quieren decir por “trastorno mental”.

· Si definimos la esquizofrenia como un trastorno al nivel de la conciencia y autoconciencia, cualquier profesional que haga uso de tales categorías debería ser capaz de lidiar preguntas tales como ¿qué es la conciencia? y ¿qué es la autoconciencia? Aunque podríamos dar una respuesta naturalista a tales preguntas, por definición, ellas pertenecen al campo de la filosofía de la mente. Junto con esto, el estudio general de las enfermedades mentales ilumina de forma particular el entendimiento de las características fundamentales de los múltiples estados mentales conscientes que los humanos gozan tales como percepciones, creencias, deseos, pensamientos, etc. por medio de la descripción fenomenológica de ciertos síntomas característicos (Zahavi, 2005a).

En consecuencia, en ambos casos podemos observar que el estudio de las enfermedades mentales propone un lugar propicio para el desarrollo de enfoques comprensivos interdisciplinarios para entender la naturaleza de la mente humana. Así, un acabado entendimiento del fenómeno no solo significaría un aporte práctico –como ya lo he mencionado– sino que también una contribución a diversas disciplinas como la psiquiatría, medicina, psicología clínica, filosofía de la mente y la fenomenología.

Finalmente, podemos observar que la comprensión de las enfermedades mentales no solo propone debates médicos y prácticos sino que también profundas discusiones filosóficas que demandan una apertura disciplinar sin la cual no se podría llegar a conclusiones plausibles (Parnas y Zahavi,

2000). Así, en las siguientes secciones intentaré explorar la relevancia filosófica del estudio de las enfermedades mentales, específicamente de la esquizofrenia, focalizando mi análisis en tres puntos principalmente: (1) la complementariedad disciplinar entre filosofía de la mente, fenomenología y psiquiatría que el entendimiento de las enfermedades demanda y la forma en que se debería entender la relación entre tales disciplinas; (2) la discusión respecto de la naturaleza y ocurrencia de ciertos síntomas esquizofrénicos; y (3) la forma en que tales síntomas desafían algunas nociones comúnmente asociadas a la conciencia humana. Con esto, planeo mostrar que el avance en el estudio de la mente está supeditado al trabajo interdisciplinario conjunto de psicólogos, médicos, psiquiatras, biólogos y filósofos, entre muchos otros.

Complementariedad disciplinar como requerimiento metodológico para la comprensión de la esquizofrenia

El término “esquizofrenia” fue propuesto por Eugene Bleuler como reemplazo para el término “demencia precoz” empleado por Emil Kraepelin para referirse a sujetos que reportaban haber vivido desconcertantes experiencias conscientes y presentaban ciertos síntomas que eran definidos como “mentales” a diferencia de los síntomas que eran definidos como puramente “físicos”. Este cambio fue motivado –principalmente– por dos razones: (1) El término “demencia” hace referencia a un profundo deterioro de las diversas funciones mentales humanas, y en los casos observados por Bleuler, tal deterioro no siempre estaba presente. Es más, Kraepelin (1896) describe la mente esquizofrénica como una “orquesta sin un director” donde las funciones mentales funcionan por separado sin unidad ni coherencia entre ellas. Por lo tanto, Bleuler concluye que los síntomas reportados no eran consistentes con la definición comúnmente asociada a las “demencias”. (2) Los síntomas observados por Bleuler y Kraepelin –actualmente denominados síntomas “esquizofrénicos”– si bien pueden aparecer en la adolescencia (Park y Nasrallah, 2014) no siempre se manifiestan en esta etapa del desarrollo humano, por lo que el término “precoz” tampoco parece hacerle justicia al fenómeno. Finalmente, Bleuer propone el término “esquizofrenia” para describir un fenómeno

caracterizado por la separación entre las diversas funciones psíquicas tales como el pensamiento, percepción, memoria, etc., lo cual parece ser consistente con algunas de las visiones críticas en la actualidad (Bental, 2003; Sass, 2013).

Pues bien, sea cual sea el caso, cuando pensamos seriamente sobre la esquizofrenia –tal como hemos visto en la introducción– es claro que su comprensión demanda un enfoque interdisciplinario que implica la integración de diversas áreas tales como la psicopatología, filosofía de la mente, fenomenología y la psicología empírica. Sin embargo, en este momento parece razonable preguntarse ¿cómo deberíamos comprender la relación entre todas estas diferentes áreas? ¿Cuál es el rol de cada disciplina en el entendimiento de este fenómeno? Solo contestando estas preguntas podremos definir de manera clara las formas más beneficiosas de investigar este trastorno y la naturaleza de la mente humana en general.

Primero, los estudios en psicopatología y psiquiatría parecen ser cruciales para determinar los méritos que las diversas teorías sobre la mente en filosofía, ciencias cognitivas y fenomenología poseen (ver Gallagher, 2012). Parece plausible indicar –cómo lo sugieren Aristóteles y Heidegger, entre otros– que la filosofía es la disciplina encargada de proporcionar descripciones y una comprensión detallada de los aspectos fundamentales de la existencia y la realidad. En el caso específico de la filosofía de la mente, el foco principal es la comprensión de los aspectos fundamentales de los fenómenos mentales. A su vez, la fenomenología intenta comprender los aspectos fundamentales de la experiencia humana. Por lo tanto, cualquier teoría que intente ser aceptada en tales disciplinas debiese lidiar de forma plausible ante el desafío propuesto por diversos contra-ejemplos empíricos provenientes desde la psicopatología.

Por ejemplo, algunos podrían indicar que la “sensación de agencia” (*sense of agency*) es una propiedad característica de nuestras experiencias conscientes. Sin embargo, investigaciones psiquiátricas sobre alucinaciones auditivas o delirios de inserción de pensamiento presentan evidencia empírica en contra de este postulado. En ambos casos, los pacientes poseen experiencias conscientes en primera persona, sin que estas estén acompañadas por la sensación fenomenológica de agencia (Bayne, 2008). En algunos casos específicos, los pacientes experimentan ciertos estados mentales como si fueran generados por agentes externos (Mullins y Spence,

2003; López-Silva, 2018). Así, cuando nos situamos en la discusión sobre la naturaleza de la consciencia, experiencia, auto-consciencia o identidad personal, entre muchas otras, los ejemplos provenientes desde la psicopatología clínica ofrecen un excelente medio para probar la plausibilidad de las teorías en filosofía de la mente y fenomenología.

En segundo lugar, Nagel (1974) acertadamente indica que si un investigador desea ofrecer explicaciones plausibles para cierta entidad –en filosofía o en cualquier disciplina–, este debe ofrecer primero que todo una descripción acabada y precisa del fenómeno que quiere explicar. En otras palabras, primero debemos distinguir el fenómeno principal de otros fenómenos secundarios asociados, con el fin de construir explicaciones precisas y exclusivas para este; lo cual parece ser sumamente crucial para el campo de mente y para disciplinas como la psiquiatría, psicopatología y la filosofía de la mente por ejemplo. Es en este momento donde se hace claro el rol de la fenomenología y la forma en que esta complementa el estudio de la mente dado que las descripciones detalladas que ofrece sobre la estructura de los fenómenos conscientes constituiría un marco metodológico unificador inicial para la investigación de los fenómenos mentales (Parnas y Handest, 2003). Cuando lidiamos con entidades tales como percepciones, pensamientos, creencias, memorias, entre otros; lidiamos con entidades que parecen tener un cierto componente cualitativo que las hace subjetivas[3]. Así, la fenomenología parece ofrecer un excelente mapeo inicial de tales entidades con valor experiencial y así, un primer paso unificador para el estudio de tales fenómenos. De forma importante, toda teoría en ciencias cognitivas, psiquiatría y filosofía de la mente debería ser capaz de explicar la fenomenología que presentan ciertos fenómenos mentales. De esta forma, podemos observar una clara complementariedad metodológica entre psiquiatría, filosofía de la mente, ciencias cognitivas, entre otras disciplinas y la fenomenología, dado que las detalladas descripciones que este ofrece respecto de la naturaleza de los fenómenos mentales humanos, proveen un punto de partida sistemático y firme para cualquier intento explicativo de las experiencia psicopatológicas.

Tercero, es sabido que la psiquiatría intentó desde sus orígenes modernos estar comprometida con una visión a-teórica y puramente empírica de los trastornos mentales, lo cual –parece– constituir una total ilusión (Bentall, 2003). Problemáticamente, esto produjo la reproducción de ciertos términos

teóricos –por parte de la comunidad médica– sin su respectiva comprensión sistemática. Términos tales como "identidad personal", "*self*", "sentido de la propia continuidad", "sensación de agencia", "mismidad", "conciencia", "estado mental", entre muchos otros, son regularmente usados en psiquiatría. Sin embargo, su definición pertenece al campo de la filosofía y las ciencias sociales, y muchas veces el uso de estos términos está lejos de ser realmente comprendido por algunas comunidades que los emplean. Acá, la discusión en filosofía y ciencias sociales determinaría la posibilidad de la existencia de un marco conceptual unificador para la investigación en el campo de la mente. Tal como ya hemos visto en la sección anterior, cuando definimos la esquizofrenia como un trastorno "mental" debemos lidiar con la pregunta sobre lo "mental". Lo mismo se aplica cuando la definimos como un trastorno de la "conciencia", donde el término "conciencia" está lejos de tener una definición clara y final.

Finalmente, ante la pregunta sobre cómo deberíamos entender la relación entre las diversas áreas relacionadas con la comprensión de la mente y sus trastornos, debiésemos contestar que, para que cada una de ellas avance, se deben necesariamente integrar los análisis de las otras áreas (Parnas y Zahavi, 2000). Así, por ejemplo, la psiquiatría debiese incluir profundos análisis filosóficos de los términos que ella emplea, y así las teorizaciones en filosofía de la mente deberían estar sujetas a la evidencia empírica provista por los estudios en psiquiatría y psicopatología. Así, esta complementariedad interdisciplinar no parecer ser una mera opción sino que requisito insoslayable para el correcto entendimiento de los fenómenos mentales de forma plausible, incluyendo la esquizofrenia.

Esquizofrenia y filosofía: discusiones abiertas

En la sección anterior he intentado explorar brevemente la forma en que debiésemos entender la relación de complementariedad entre las diversas disciplinas teóricas y aplicadas que intentan lidiar con los fenómenos mentales. Sin embargo, esta complementariedad metodológica no es el único elemento que se deriva del estudio de los fenómenos mentales y sus respectivos trastornos. Desde un punto de vista teórico, el entendimiento de la estructura intencional de ciertos síntomas de la esquizofrenia propone un

desafío para ciertas teorías actuales sobre la consciencia y la autoconsciencia humana (Metzinger, 2003; 2005 entre otros). Además, el análisis de tales síntomas propone preguntas generales sobre la naturaleza de las creencias humanas en tanto actitudes proposicionales, y preguntas específicas respecto de la naturaleza y origen de los estados alucinatorios que caracterizan ciertos trastornos mentales (Campbell; 1999; 2002; Stephens y Graham, 2000; Sollberger, 2013).

En esta sección, primero introduciré al lector al debate filosófico general sobre el origen y naturaleza de los delirios de inserción de pensamiento, un síntoma clave en el diagnóstico de la esquizofrenia. Luego, intentaré clarificar la forma en que este síntoma esquizofrénico provee evidencia en contra de la teoría fenomenológica de la autoconciencia defendida por Dan Zahavi (1999; 2005a; 2011). Así, este análisis servirá como una ejemplificación de la relación interdisciplinaria necesaria para entender los fenómenos mentales aludida en la sección anterior, demostrando que la examinación filosófica de diversas experiencias psicopatológicas –como las que caracterizan la esquizofrenia–, no solo informa, sino que direcciona el estudio de las características fundamentales de la experiencia consciente.

El enfoque presentacional de la autoconsciencia humana

En términos generales, el enfoque presentacional es un enfoque fenomenológico que propone que todo episodio de consciencia fenoménica –esto es, episodios experienciales (Block, 1995)– lleva necesariamente consigo un grado mínimo de autoconsciencia (*minimal self-awareness*), de forma tal que este sí-mismo del cual somos consciente en nuestro experienciar es intrínseco a la estructura intencional de toda experiencia (Zahavi, 2005a; Grünbaum y Zahavi, 2011). Tal como Zahavi (2011) indica: "[*self-awareness*] *is not something added to the experience, an additional mental state, but rather an intrinsic feature of the experience*"[4] (Zahavi, 2011, pp. 56-57). Por lo tanto, en este enfoque, el sujeto (que es parte de la estructura de toda experiencia) no es algo que se le "suma" o algo "diferente" de la experiencia misma como lo postulan las teorías representacionales de segundo orden de la conciencia (Rosenthal, 1997; Carruthers, 1996; entre otros). En la enfoque presentacional, el sí-mismo es

un elemento de la forma natural en que cualquier experiencia “es dada” a la conciencia.

Siguiendo una distinción originalmente introducida por Husserl en *Ideen I*, Dan Zahavi (2005a) indica que toda experiencia consciente es “presentacional” i.e. todas “presentan” objetos teniendo diversas propiedades. Esta estructura fenoménica presentacional poseería dos dimensiones inseparables fundamentales, por una parte la dimensión intencional externa, que refiere al objeto mismo de la experiencia y sus características. Por ejemplo: una copa de vino blanco y su sabor, color, etc. Por otra parte, la dimensión intencional interna: dada por la modalidad experiencial por medio de la cual un cierto objeto es dado a la consciencia, por ejemplo, recordar una copa de vino, pensar en una copa de vino, ver una copa de vino, etc. En todos estos casos el objeto intencional es el mismo (la copa de vino), pero ese objeto puede ser dado en una modalidad experiencial diferente.

Un punto crucial para esta teoría, es que Zahavi (2005a) sostiene que todas estas modalidades experienciales comparten una propiedad fenoménica en común: todas ellas poseen una “sensación de propiedad” y la existencia de tal propiedad será el principal argumento para la existencia de este sí-mismo intrínseco a la estructura de toda experiencia en este enfoque. El término empleado originalmente es *sense of mineness* (ver Zahavi, 2005a, p. 124). Sin embargo, en algunas publicaciones en español (ver por ejemplo la traducción de Dörr e Irarrázaval del EASE: Examen de Anomalías Subjetivas de la Experiencia de Parnas *et al.,* 2005; ver también Rossi-Monti, 2008) el término ha sido incorrectamente traducido como “mismidad”. Sin embargo, mismidad en inglés es *sameness* y refiere a la propiedad de un ente de ser el mismo a través del tiempo. Por otra parte, el *sense of mineness*, refiere al hecho de que mis experiencias conscientes son dadas como mías o al sentido de que mis experiencias me pertenecen a mí. Por esta razón, he decidido emplear mi propia traducción del término i.e. “sensación de propiedad”.

Ahora volviendo a la discusión, el argumento de Zahavi es principalmente fenomenológico. La sensación de propiedad refiere al hecho que toda experiencia es siempre “dada” para un sujeto, todas mis experiencias conscientes son dadas como “mías”, en el sentido de que cada una de ellas me pertenece a mi (Zahavi y Parnas, 1998; Zahavi, 2002;

Zahavi 2005a; Gallagher, 2012). La idea principal es que uno tiene un tipo de acceso especial a sus experiencias que es diferente del tipo de acceso que los demás tienen a ellas. Mientras los otros acceden a mis experiencias desde el punto de vista de la tercera persona, yo accedo en primera persona y este modo de acceso es caracterizado por la existencia de esta sensación de propiedad.

Ahora bien, la premisa fundamental es que la sensación de propiedad revela al sujeto de cada experiencia como un elemento intrínsecamente contenido en su estructura intencional. Así, uno deviene consciente de ser el sujeto de cierta experiencia en tanto la experiencia es dada como "mí experiencia". Por lo tanto, Zahavi (2005a) entiende la sensación de propiedad como una condición necesaria para la existencia de la auto-consciencia humana sin tener que recurrir a elementos adicionales a la experiencia misma, como lo hacen sus contendores. Un punto crucial para notar es que Zahavi (2005a) indica que la sensación de propiedad (*sense of mineness*) es intrínseca a todas las modalidades experienciales, y por lo tanto, todas las experiencias conscientes conllevan conciencia del sujeto que las protagoniza (*self-awareness*). Finalmente, para Zahavi, toda experiencia necesariamente posee una sensación de propiedad y es en virtud de esta que cada experiencia lleva consigo auto-consciencia.

Auto-conciencia y esquizofrenia

Pues bien, el enfoque presentacional de Zahavi parece ser desafiado directamente por la fenomenología asociada a uno de los síntomas que ha guiado el diagnóstico de la esquizofrenia durante los últimos 40 años i.e. los delirios de inserción de pensamiento (Mullins y Spence, 2003). Pacientes que sufren de delirios de inserción de pensamiento reportan que cierta idea o pensamiento ha sido "introducido" o "puesta" en sus mentes o en "su cabeza" (Mellor, 1970; Zahavi, 2005a; Roessler, 2013). El siguiente, es un reporte clásico de este síntoma: "*Thoughts are put into my mind like 'Kill God'. It's just like my mind working, but it isn't. They come from this chap, Chris. They're his thoughts*"[5] (Frith, 1992, p. 66).

Acá, el paciente indica que "Chris" (un presentador de la televisión inglesa) ha puesto una cierta idea en su cabeza ("Kill God"). Existe cierto

consenso en la comunidad psiquiátrica y filosófica al indicar que la principal característica fenomenológica de este delirio es que los pacientes no experimentan una sensación de propiedad asociada al pensamiento que ha sido insertado (Martin y Pacherie, 2013). Los pacientes indican cosas como “esta idea no es mía”, “esta idea no me pertenece”, o que “¡este pensamiento no es mío!”. Luego, el elemento más relevante para nuestro análisis es que este síntoma de la esquizofrenia reflejaría una anomalía en la forma en que los pacientes experimentan sus pensamientos que es contraria a la expuesta por el enfoque presentacional. La ocurrencia de los delirios de inserción de pensamiento hace plausible concluir que en estos casos, la condición para la existencia de la auto-consciencia propuesta por Zahavi no se cumple dado que es posible observar casos en los cuales un episodio experiencial (el pensamiento) no posee una sensación de propiedad. Así, si es posible concebir casos reales en los cuáles ciertas experiencias consientes no son dadas a la consciencia con tal propiedad, entonces uno no debería aceptar el argumento entregado por el enfoque presentacional.

Tal conclusión parece ser consistente con uno de los aspectos positivos del delirio denominado “atribución externa” (Campbell, 1999; Stephens y Graham, 2000; entre otros). Este aspecto refiere al hecho de que los pacientes comúnmente identifican al agente que ha insertado la idea o pensamiento en sus cabezas (Saks, 2008). Este aspecto es comúnmente explicado como una forma patológica de lidiar con la inicial ausencia de sensación de propiedad (algo así como una explicación patológica de la experiencia anómala), lo cual afirma nuestra conclusión inicial. Si bien la discusión es más compleja que esta breve caricatura, lo que es claro es que el estudio de los síntomas de la esquizofrenia contribuye activamente a la refutación o aceptación de ciertas teorías que intentan lidiar con la naturaleza de la mente humana y sus características[6]. Ahora bien, no es mi propósito resolver la discusión en este trabajo; sin embargo, es interesante observar que el desafío propuesto por los delirios de inserción de pensamiento ha producido un debate en fenomenología, filosofía de la mente y psiquiatría filosófica que está actualmente abierto (Martin y Pacherie, 2013; Roessler, 2013, entre otros).

La naturaleza de la auto-consciencia

Tal como hemos visto en la sección anterior, los delirios de inserción de pensamiento desafían diversas teorías sobre la autoconsciencia humana (ver Metzinger, 2003; Zahavi, 2005a y Billon, 2013). Sin embargo, tal discusión se mantiene abierta. Dan Zahavi ha elaborado al menos dos estrategias para poder lidiar con el desafío propuesto por este delirio; analicemos ahora brevemente solo una de ellas.

Zahavi (2005a; 2005b) y Grünbaum y Zahavi (2013) han reinterpretado los casos de delirios de inserción de pensamiento indicando que estos representarían una alteración en la sensación de agencia, pero no en la sensación de propiedad. En tales casos, este último quedaría intacto. Zahavi (2005b) señala:

> *When schizophrenics assert that their thoughts are not their own, they do not mean that they themselves are not having the thoughts, but rather that someone else has inserted them and that they themselves are not responsible for generating them. Thus, rather than involving a lack of a sense of ownership, passivity phenomena like thought insertions involve a lack of a sense of authorship (or self-agency) and a misattribution of agency to someone or something else*[7] (p. 6).

Como podemos ver, Zahavi intenta salvar su enfoque mediante la reinterpretación de la fenomenología de los delirios de inserción de pensamiento. Sin embargo, esta estrategia parece tener diversos problemas a nivel conceptual y fenomenológico.

El primer problema con esta estrategia es conceptual. La sensación de agencia es comúnmente asociada a "causalidad intencional", esto es, el hecho de que los sujetos poseen la intención de pensar un cierto pensamiento o idea que es previa a la ocurrencia misma de tal pensamiento (Wegner y Sparrow, 2004). Así, cuando el sujeto no logra identificar tal intención –por distintas razones– este no retendría una sensación de agencia. El problema es que una intención "I" para pensar "P" en sí misma necesitaría estar precedida por otra intención "I*" dado que "I" es un tipo de pensamiento. Esto es un problema, dado que "I*" también es un tipo de pensamiento y todo pensamiento necesita una intención de lo precede, "I*" necesitaría una intención previa denominada "I**" y así *ad infinitum*. Luego, si todo pensamiento es explicado bajo la base de una intención previa a pensar tal pensamiento, la distinción propuesta por Zahavi remite al infinito cada vez que un agente piensa[8].

El segundo problema de esta estrategia es fenomenológico. La distinción

no lograría discriminar adecuadamente entre la fenomenología asociada a los delirios de inserción de pensamiento y la fenomenología asociada a otras experiencias que no son patológicas pero que también se explican por la retención de una sensación de propiedad y la alteración –no patológica– de la sensación de agencia. Por ejemplo, tomemos el caso de los "pensamientos repentinos" i.e. aquellos pensamientos o ideas que aparecen repentina y sorpresivamente en nuestro flujo de la consciencia sin que podamos reconocer ninguna intención de pensarlos. Pues bien, este tipo de actividad mental también puede ser descrita como un caso en el cual uno no reconoce una sensación de agencia pero si conserva una sensación de propiedad (algo así como "no tuve la intención de pensar esta idea, pero a pesar de todo esta idea es experimentada como mía"). De forma problemática para Zahavi, la fenomenología asociada a este tipo de actividad mental no está ni remotamente cerca de la fenomenología asociada a los delirios de inserción de pensamiento (Cermolacce, Naudin y Parnas, 2008). La misma objeción puede ser estructurada a partir del caso de los pensamientos obsesivos. Un pensamiento obsesivo aparece en nuestra consciencia sin que podamos reconocer una clara intención de pensarlo, sin embargo, tales pensamientos son reconocidos como nuestros –sensación de propiedad– a pesar de su contenido molesto o poco placentero. Igualmente, la fenomenología de un pensamiento obsesivo, dista mucho de la fenomenología reportada por los pacientes que sufren de delirios de inserción de pensamiento.

Una consecuencia derivada de este último punto, es que la reinterpretación antojadiza de la fenomenología de los delirios de inserción de pensamiento en la cual incurre Zahavi provoca que el síntoma pierda todas sus características patológicas. Cuando el autor explica de la misma manera fenómenos normales y patológicos, esté niega completamente el tilde negativo que permea todas las descripciones de los pacientes esquizofrénicos. Tales pacientes continuamente demuestran un conflicto derivado de la ocurrencia de experiencias conscientes que les son difíciles de aceptar y con las cuales no pueden lidiar muchas veces. Este fenómeno de "despatologización" que caracteriza la estrategia del autor, es muestra de que esta es *ad hoc* y de que, por lo tanto, no debiésemos aceptar tal respuesta.

Tal como podemos ver, tal discusión está lejos de estar resuelta. Hasta

este punto, he intentado solamente introducir al lector al actual estado del debate sobre la comprensión de la relación entre la estructura de las experiencias conscientes, autoconciencia y esquizofrenia. A su vez, esto muestra que el mejor camino que el estudio de la mente puede seguir es aquel que se caracterice por el diálogo interdisciplinario.

La naturaleza de los delirios: una introducción al debate

Actualmente, múltiples teorías rivales intentan explicar cómo es posible que los pacientes que sufren de delirios de inserción de pensamiento experimenten el tipo de fenómenos conscientes que parecen indicar i.e. que un agente externo introduzca ciertas ideas o pensamiento en sus cabezas. Por ejemplo, en un artículo seminal, Campbell (1999) indica que los delirios de inserción de pensamiento consisten en la ruptura de la unidad de las diversas propiedades fenoménicas de la experiencia consciente. Así, estas se separarían en ciertos estados psicopatológicos. El autor indica que todos los estados mentales son dados para un sujeto con lo que se denomina: "sensación de pertenencia", que refiere a calidad de que cierto estado mental es "mío", tal como lo revisamos en la sección anterior. Sin embargo, existirían dos formas de pertenencia que se podrían separar en la ocurrencia de los delirios de inserción de pensamiento. La primera es la pertenencia propiamente tal, que estaría dada por el mero hecho de que cierto estado mental está dado en mi flujo de consciencia (Grünbaum y Zahavi, 2013). Esta idea ha sido actualmente debatida y posee diversos detractores (Martin y Pacherie, 2013; López-Silva, 2013b). La segunda forma de pertenencia es la sensación de agencia, que refiere a la sensación de ser el autor de nuestros propios estados mentales, esto es, la propiedad asociada a un estado mental de que yo soy quien lo ha iniciado (en el caso de cierta acción) o creado (en el caso de un pensamiento) (Bayne, 2008; Gallagher, 2001; 2004; 2012).

Una de las aristas del debate filosófico consiste en explicar el origen y naturaleza de la sensación de agencia, que sería uno de los rasgos que se perdería en los delirios de inserción de pensamiento[9]. Así, parecen existir al menos dos enfoques para poder explicar la pérdida de la sensación de agencia: (1) El enfoque "psicológico" indica que la ausencia de sensación

de agencia debería ser explicada como una reacción racional del paciente ante ideas o pensamientos de naturaleza extraña y calidad cualitativa negativa. Así, la falta de consistencia y sincronización de tales ideas con las creencias cotidianas del paciente, sería combatida con una negación de la agencia de tal idea (Campbell, 1999; Vosgerau y Newen, 2007). (2) El enfoque de "monitoreo cognitivo" atribuye la ausencia de sensación de agencia característica de los delirios de inserción de pensamiento a una falla en las habilidades de auto-monitoreo del paciente (Frith, 1992). Este enfoque asume que los humanos continuamente monitoreamos nuestras acciones motores y mentales. Así, en el caso de un pensamiento, uno monitorearía la ocurrencia de tal pensamiento mediante la contrastación de este con una previa intención a pensar tal pensamiento. De alguna manera, este proceso sería interrumpido en esquizofrenia con el resultado de la carencia de la sensación de agencia (ver Gallagher, 2004).

Ambos enfoques han sido continuamente criticados (Gibbs, 2000; Vosgerau y Newen, 2007; Bortolotti y Broome, 2008; Martin y Pacherie, 2013) y aunque la literatura en el debate sigue creciendo (Billon, 2013; Roessler, 2013), este se halla lejos de estar zanjado.

Conclusiones: oportunidades y desafíos en el estudio de la mente

Los síntomas característicos de la esquizofrenia no implican la mera ocurrencia de experiencias conscientes con contenidos anormales, sino que también parecen conllevar profundas anomalías en la estructura de la conciencia y autoconciencia humana. Así, la comprensión de las enfermedades mentales no solo propone debates médicos sino que también profundas discusiones filosóficas acerca de la naturaleza de los fenómenos mentales. A lo largo de este capítulo he intentado explorar la relevancia filosófica que posee el estudio y comprensión de los trastornos mentales, especialmente, la comprensión de la esquizofrenia. El análisis estuvo centrado en la complementariedad disciplinar entre filosofía de la mente, fenomenología y psiquiatría que el entendimiento de la esquizofrenia parece demandar como condición necesaria para llegar a conclusiones plausibles en el estudio de la mente. Luego de mapear la forma en que estas distintas disciplinas se interrelacionan, el análisis se centró en la discusión general

respecto de la naturaleza de los delirios de inserción de pensamiento y en la forma en que este síntoma desafía algunas nociones comúnmente asociadas a la conciencia humana. Ahora bien, los diversos puntos analizados en este capítulo proponen una oportunidad invaluable de cooperación multidisciplinar. El estudio de las alucinaciones, los delirios, el diagnóstico de las enfermedades mentales, la agencia humana, la consciencia, los trastornos afectivos, entre muchos otros conceptos, parecen demandar un diálogo entre diversas disciplinas. Dado esto, es responsabilidad de los actuales profesionales trabajando en las diversas disciplinas que intentan lidiar con la mente humana el de generar instancias de trabajo conjunto con el fin de complementar sus conclusiones y así, mejorar la calidad de cada una de sus contribuciones a sus respectivas disciplinas.

Agradecimientos

Quisiera agradecer al Dr. Joel Smith (University of Manchester), a la audiencia del seminario “Enfoques Interdisciplinario en Filosofía de la Mente y Psicología” (Universidad de Chile, Diciembre, 2013) y a la audiencia del “Conversatorio de las Ciencias Sociales en Arica y Tacna” (Universidad de Tarapacá y Universidad Privada de Tacna) por sus útiles comentarios en algunos de los puntos expuestos en este capítulo. La escritura de este artículo fue financiada por el proyecto Fondecyt Nº 11160544 “La arquitectura agencial del pensamiento humano” de la Comisión Nacional de Investigación Científica y Tecnológica (Conicyt) del Gobierno de Chile.

Referencias

Armstrong, D.M. (1968). *A Materialist Theory of the Mind*, London: Routledge.

Bayne, T. (2008). The phenomenology of agency. Philosophy Compass, 3(1), pp. 182-202.

Bayne, T. y Montague, M. (2012). Cognitive Phenomenology. Oxford: OUP.

Bentall, R. (2003). Madness Explained. UK: Pinguin.

Billon, A. (2013). Does consciousness entail subjectivity? The puzzle of thought insertion. Philosophical Psychology, 26(2), pp. 291-314.

Boghossian, P. (2006). *Fear of knowledge: Against relativism and constructivism*. UK: Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences,* 18, pp. 227-247.

Bortolotti, L., y Broome, M. (2009). A role for ownership and authorship in the analysis of thought insertion. Phenomenology and the Cognitive Sciences, 8, pp. 205-224.

Campbell, J. (2002). The ownership of thoughts. Philosophy, Psychiatry, y Psychology, 9(1), pp. 35-39.

—— (1999). Schizophrenia, the space of reasons, and thinking as motor process. The Monist, 84, pp. 609-625.

Carruthers, P., 1996. *Language, Thoughts and Consciousness. An Essay in Philosophical Psychology*, Cambridge: Cambridge University Press.

Cermolacce, M., Naudin, J. y Parnas, J. (2007). The '"minimal self'" in psychopathology: Re-examining the self-disorders in the schizophrenia spectrum. Consciousness and Cognition 16 (2007), pp. 703-714.

Cooper, R. (2013). Natural Kinds. En K. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini y T. Thornton (Eds.), The Oxford Handbook of Philosophy and Psychiatry. Oxford: OUP.

—— (2007). Psychiatry and Philosophy of Science. London, Acumen Publishing Ltd.

Crane, T. (1996), *The Mechanical Mind*: *A Philosophical Introduction to Minds, Machines and Mental Representation*, London: Penguin.

Frith, C.D. (1992). The cognitive neuropsychology of schizophrenia. Hoeve: Lawrennce Erlbaum Associates.

Fulford, K. Davies, M., Gipps, R., Graham, G., Sadler, J., Stanghellini, G. y Thornton, T. (Eds.) (2013). The Oxford Handbook of Philosophy and Psychiatry. Oxford: OUP.

Gallagher, S. (2012). Phenomenology. U.K.: Macmillan.

—— (2004). Neurocognitive models of schizophrenia. A neurophenomenological critique. Psychopathology, 37, pp. 8-19.

—— (2000), "Philosophical Conceptions of the Self: Implications for Cognitive Science", Trends in Cognitive Sciences 4, pp. 14-21.

Gergen, K. (1991). *The Saturated Self*. New York: Basic Books.

Gibbs, P. (2000). Thought Insertion and the Inseparability Thesis. Philosophy, Psychiatry, y Psychology, Volume 7, Nº 3, September 2000, pp. 195-202.

Grünbaum, T. y Zahavi, D. (2013). Varieties of self-awareness. In K. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini y T. Thornton (Eds.), The Oxford Handbook of Philosophy and Psychiatry, pp. 221-239. Oxford: OUP.

Hacking, I., (1999). *The Social Construction of What?* USA: Harvard University Press.

Jackson, F., (1986). What Mary Didn't Know. *Journal of Philosophy*, LXXXIII, pp. 291-295.

Kraepelin, E. (1896). Psychiatrie (4th edn). Leipzing: J. A. Barth.

Lakatos (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers Volume*

1. Cambridge: Cambridge University Press.

López-Silva, P. (2018). Mapping the Psychotic Mind. *Psychiatric Quarterly* [Online First], pp. 1-12.

—— (2017). Me and I are not friends, just Acquaintances: On thought Insertion and Self-Awareness. *The Review of Philosophy y Psychology.* DOI: https://doi.org/10.1007/s13164-017-0366-z.

—— (2013a). Construcciones, Realidades y Dilemas. Una revisión filosófica al construccionismo social. *Cinta de Moebio, Revista de Epistemología de la Universidad de Chile*, 46, pp. 9-25.

—— (2013a). Autoconsciencia y la Fenomenología de los Síntomas Esquizofrénicos. Ponencia presentada en el seminario "Enfoques Interdisciplinarios en Filosofía de la Mente y Psicología" (Universidad de Chile), Santiago, Chile.

Martin, J-M. y Pacherie, E. (2013). Out of Nowhere: Thought Insertion, Ownership and Context-Integration. Consciousness and Cognition, 22(1), pp. 111-122.

Marwaha, S. Broome, M., Bebbington, P., Kuipers, E. y Freeman, D. (2013). Mood Instability and Psychosis: Analyses of British National Survey Data. Schizophrenia Bulletin, Published onlieve first: http://schizophreniabulletin.oxfordjournals.org/content/early/2013/10/25/schbul.sbt149.short. DOI: 10.1093/schbul/sbt149.

Mellor, C. S. (1970). First Rank symptoms of schizophrenia. The British journal of Psychiatry, 117, pp. 15-23.

Metzinger, T. (2006). Reply to Zahavi: The Value of Historical Scholarship. Psyche, 12(2), pp. 1-4.

—— (2003). Being no one. Cambridge: MIT Press.

Mullins, S. y Spence, S. (2003). Re-examining thought insertion. British journal of psychiatry (2003), 182, pp. 293-298.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review* 83, pp. 435-450.

Park, S. y Nasrallah, H. (2014). The varieties of anomalous self experiences in schizophrenia: Splitting of the mind at a crossroad. Schizophrenia Research. Available online form: http://www.vanderbilt.edu/parklab/Park%20and%20Nasrallah%202013.pdf.

Parnas, J., Møller, P., Kircher, T., Thalbitzer, J., Jansson, L., Handest, P. y Zahavi, D. (2005). EASE: Examination of Anomalous Self-Experience. Psychopathology, 38, pp. 236-258.

Parnas, J. y Sass, L. (2001). Self, solipsisim, and schizophrenia. Philosophy, Psychiatry, y Psychology. 8(2/3), pp. 101-120.

Parnas, J. y P. Handest (2003). Phenomenology of Anomalous Self-Experience in Early Schizophrenia. Comprehensive Psychiatry, Vol. 44, Nº 2 (March/April), pp. 121-134.

Parnas, J. y Zahavi, D. (2000). The link: philosophy-Psychopathology-Phenomenology. En D. Zahavi (Ed.) Exploring the self, pp. 1-16. Amsterdam: John Benjamins.

Pickard, H. (2009). Mental illness is indeed a myth. En M.R. Broome y L. Bortolotti (Eds.) Psychiatry as cognitive neuroscience: Philosophical perspectives. Oxford: OUP.

Poland, J. y Von Eckardt, B. (2013). Mapping the Domain of Mental Illness. En K. Fulford, M.

Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini y T. Thornton (Eds.), The Oxford Handbook of Philosophy and Psychiatry. Oxford: OUP.

Roessler, J. (2013). Thought Insertion, Self-Awareness, and Rationality. En K. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini y T. Thornton (Eds.), The Oxford Handbook of Philosophy and Psychiatry. Oxford: OUP.

Rosenthal, D. M., 1997. "A Theory of Consciousness" in N. Block, O. Flanagan and G. Güzeldere (eds.) *The Nature of Consciousness*, Cambridge, MA: MIT Press, pp. 729-753.

Rossi-Monti, M. (2008). En los límites de la identidad: El caso de los delirios extraños. Archivos de Psiquiatría, 2-4(71), pp. 80-95.

Saks, E. R. (2007). The centre cannot hold. My journey through madness. New York: Hyperion.

Sass, L. (2013). Self-disturbance and schizophrenia: Structure, specificity, pathogenesis (Current issues, New directions). Schizophrenia Research. Available online: http://www.sciencedirect.com/science/article/pii/S0920996413002727. DOI: http://dx.doi.org/10.1016/j.schres.2013.05.017.

Searle, J. (1980). Minds, Brains and Programs, *Behavioral and Brain Sciences*, 3, 417-57.

Sollberger, M. (2013). Making sense of an endorsement model of thought insertion. Available online: http://doc.rero.ch/record/31858/files/Sollberger_Making_sense_of_an_endorsement_model_of_thought-insertion_personal_version_.pdf.

Stephens, G. y Graham, G. (2000). When Self-Consciosness Breaks Down. USA: MIT Press.

Szasz, T. (1974). The Myth of Mental Illness. London, Palladin.

Vosgerau G. y Newen, A. 2007. Thoughts, Motor Actions, and the Self. Mind y Language, Vol. 22 No. 1 February 2007, pp. 22-43.

Wegner, D., y Sparrow, B. (2004). Authorship processing. In M. S. Gazzaniga (Ed.), The new cognitive neurosciences (third ed.). USA: MIT.

Werner P, Aviv A. y Borak Y. (2008). Self-stigma, self-esteem and age in persons with schizophrenia. Int Psychogeriatr. 20(1), pp. 174-187.

Zahavi, D. (2011). The experiential self: Objections and clarifications. In M. Siderits, E. Thompson y D. Zahavi (Eds.), Self, No Self?, pp. 56-78. Oxford: OUP.

—— (2005a). Subjectivity and Selfhood: Investigating the first-person perspective. Massachusetts: MIT Press.

—— (2005b). Being someone. Psyche, 11(5), 1-20.

—— (1999). Self-awareness and alterity: A phenomenological investigation. Evanston, IL: Northwestern University Press.

Zahavi, D. y Parnas, J. (1998). Phenomenal consciousness and self-awareness: a phenomenological critique of representational theory. Journal of Consciousness Studies, 5, Nº 5-6, 1998, pp. 687-705.

[1] Para una revisión crítica de las diversas consecuencias sociales derivadas del diagnóstico de la esquizofrenia ver Szasz (1974).

[2] Para un análisis crítico sistemático de esta idea ver: Hacking (1999). “The Social Construction of What?”.

[3] Para una completa revisión de las discusiones asociadas al valor cualitativo de los fenómenos mentales, ver: Bayne y Montague (2012).

[4] La conciencia de sí-mismo no es algo que se le suma a la experiencia, un estado mental adicional, sino que, más bien, una característica intrínseca a la experiencia.

[5] Algunos pensamientos son puestos dentro de mi mente como “Mata a Dios”. Es como que si mi mente estuviese trabajando, pero no es así. Ellos vienen de este tipo, Chris. Son sus pensamientos.

[6] Para una revisión en profundidad de mi argumento respecto de la implausibilidad de la teoría de Zahavi, revisar López-Silva (2017).

[7] Cuando los esquizofrénicos aseguran que sus pensamientos no son suyos, ellos no quieren decir que no están teniendo tales pensamientos, sino que alguien más los ha insertado y que ellos no son responsables de su producción. Así, más que envolver una pérdida de sensación de agencia, fenómenos de pasividad como la inserción de pensamiento envuelven una falta de sensación de autoría (o auto-agencia) y una atribución errónea de agencia a algo o alguien más.

[8] Esta es una crítica comúnmente asociada a la distinción entre sensación de propiedad y sensación de agencia en los casos de delirios de inserción de pensamiento (ver Bortolotti y Broome, 2008) y que Zahavi parece reproducir en su estrategia para salvar su enfoque.

[9] Es importante notar que esta interpretación del fenómeno posee múltiples problemas (Bortolotti y Broome, 2008; Martin y Pacherie, 2013). He centrado mi análisis en tal distinción, solamente para introducir al lector a los aspectos generales de la discusión.

CAPÍTULO XI

El problema de la explicación de los delirios: en defensa de un enfoque multifactorial

Emilia Vilatta

La dirección de la explicación causal

Los delirios han sido considerados como el sello patognomónico de la locura, siendo un síntoma crucial de la psicosis (en particular de la esquizofrenia) y una característica prominente de otras condiciones psiquiátricas y neurológicas. Si bien hay mucha controversia respecto a cómo definirlos, en particular porque ninguna de las de definiciones actuales parece ofrecer criterios suficientes y necesarios (Miyazono y Bortolotti, 2015), la definición provista por el *Manual diagnóstico y estadístico de los trastornos mentales* (DSM V, 2013, p. 824) sostiene que son creencias falsas basadas en deducciones incorrectas sobre la realidad externa que se mantienen firmemente a pesar de lo que el resto de las personas creen y a pesar de las pruebas evidentes e indiscutibles de lo contrario. Tanto la definición de los delirios como su explicación han generado diversos debates filosóficos y científicos.

Respecto al desarrollo de un modelo explicativo sobre los delirios, un acuerdo que se comparte actualmente en la literatura es que habría una experiencia anómala que estaría involucrada en la generación del delirio (Bayne y Pacherie, 2004; Campbell, 2001; Davies *et al.*, 2001; Ellis y Young, 1990; Gerrans, 1999; Maher, 1974; Maher, 2001; McKay, Langdon y Coltheart, 2009). Sin embargo, existen dos tipos de enfoques respecto a cuál es la dirección de la relación causal entre la experiencia y la creencia en la formación del delirio: los enfoques "*top-down*" y los enfoques "*bottom-up*". Los primeros consideran que la dirección de la relación causal es de arriba hacia abajo, esto es, desde ciertas creencias fundamentales a la

experiencia perceptual. Cabe aclarar que, bajo este enfoque, sería un mal funcionamiento orgánico lo que generaría una alteración en ciertas creencias fundamentales, lo cual, a su vez, produciría experiencias perceptuales anómalas. Como consecuencia, el sujeto terminaría generando creencias delirantes para dar cuenta de aquellas experiencias inusuales.

Por el contrario, los defensores del segundo enfoque consideran que la dirección es de abajo hacia arriba, es decir, que va desde la experiencia hacia la formación de la creencia. Habría también un mal funcionamiento orgánico, pero este generaría una alteración en la experiencia perceptual y, como resultado, el sujeto arribaría a creencias delirantes, o bien en un intento por explicar tales experiencias, o bien porque la experiencia ya comprendería el contenido mismo del delirio[1]. Mientras que los modelos *"bottom-up"* han sido denominados enfoques empiristas, por hacer énfasis en la experiencia, los modelos "*top-down"* han sido denominados racionalistas, por poner el foco en la alteración de ciertas creencias fundamentales.

El enfoque *"top-down"*

El enfoque "*top-down*" ha sido particularmente propuesto para los delirios monotemáticos como el delirio de Cotard y el de Capgras. La propuesta más contundente ha sido desarrollada por Campbell (2001) quien revierte la dirección de la explicación causal de los enfoques "*bottom-up"*, atribuyendo directamente la causa del delirio a un disturbio en ciertas creencias fundamentales del sujeto. Este enfoque considera que serían factores orgánicos, aunque no se brinda precisión respecto a cuáles serían estos factores ni a cómo los mismos generarían dicha alteración, los que producirían un disturbio en las creencias fundamentales. Este disturbio alteraría la experiencia perceptual y se generarían, como consecuencia, las creencias delirantes.

Al considerar los casos de Capgras[2], Campbell (2001) se ocupa de argumentar por qué habría en ellos un disturbio de tipo "*top-down"*. A su juicio, para explicar cómo el sujeto delirante puede llegar a tener una experiencia perceptual con un contenido delirante tal como "aquella persona no es mi cónyuge, sino un impostor" debemos aceptar que hay

algún tipo de penetración cognitiva de tipo "*top-down*" ("*top-down loading*"), desde ciertas creencias fundamentales hacia la percepción. Asimismo, además de defender la etiología "*top-down*" de los delirios, Campbell (2001) se ocupa también de cuestionar que los sujetos delirantes estén dando su significado usual a aquellos términos mediante los cuales manifiestan sus delirios.

Partiendo de la asunción de que la racionalidad es una precondición para la atribución de estados mentales, Campbell (2001), plantea el problema de atribuir irracionalidad preguntándose si realmente el sujeto conoce los significados de las palabras que está usando, pues en la medida en que esta comprensión provee la habilidad para razonar correctamente utilizando dichos términos, ¿cómo es que sería posible razonar tan erróneamente? (p. 90). A criterio de Campbell, lo que pone de manifiesto el sujeto delirante al expresar sus creencias es que no está comprendiendo adecuadamente los términos que está utilizando, o que les está dando un significado diferente del nuestro. Como ejemplo de ello, refiere a los pacientes con síndrome de Cotard, quienes afirman que están muertos a pesar de que, por ejemplo, están hablando. Frente a esto se pregunta: "¿Cómo se puede decir realmente que los pacientes siguen teniendo conocimiento de los significados de sus declaraciones cuando están usando las palabras en una manera tan desviada?" (p. 91) y afirma que "no podemos hallar ninguna relación entre el uso que hace el paciente de las palabras y su conocimiento de las referencias de las mismas" (p. 95). En síntesis: para Campbell (2001), la clave para comprender los disturbios de los sujetos delirantes está en el uso divergente que estos hacen de los términos con los cuales expresan su delirio, y no en su experiencia anómala, tal como señalan los empiristas. Con el propósito de desarrollar esta idea, el autor establece un paralelo entre el estatus que el sujeto delirante asigna a las proposiciones que expresan el delirio y el status particular que Wittgenstein ha señalado que tienen las denominadas proposiciones marco ("*framework propositions")*.

Brevemente, las proposiciones marco expresarían un tipo particular de creencias que tendrían un estatus epistémico especial, muy diferente al de las creencias estándar. En *On Certainty* (1969), Wittgenstein propone como algunos ejemplos de estas: "La tierra es muy antigua" o "nunca he puesto los pies sobre la luna". Dichas creencias ni están justificadas, ni son justificables, pues son "certezas básicas" ("*bedrock beliefs*") que hacen

posible cualquier tipo justificación (§ 253). Asimismo, no se puede decir que son verdaderas o falsas, e incluso poner en duda las mismas sería caer en un sin-sentido. Al respecto, la idea de Campbell es que las creencias delirantes tendrían el estatus epistémico que Wittgenstein le asigna a las proposiciones marco. Esto es, serían creencias tratadas como presuposiciones de fondo, necesarias para chequear la corrección de cualquier otra proposición y no estarían en sí mismas, en ningún sentido ordinario, sujetas a algún tipo de escrutinio empírico. Así pues, de acuerdo al análisis de Campbell (2001), lo primero que encontraríamos en el caso de los sujetos delirantes es una diferencia en las proposiciones marco que estos sostienen –diferentes de aquellas proposiciones marco a las que adherirían los sujetos no delirantes– y esto es lo que explicaría la experiencia anómala que padecen. De este modo, la experiencia perceptual anómala sería una consecuencia de una diferencia más fundamental, situada en el nivel en de las proposiciones marco. A criterio de Campbell, un cambio en estas proposiciones conllevaría un cambio en los significados de los términos usados, pues "el sujeto que se mueve desde un conjunto de estas proposiciones hacia otro, desestabilizaría los significados de los términos usados y los nuevos significados no serían siquiera conmensurables con los viejos significados; no habría traducción de un marco a otro" (p. 98). Esto sería, a su criterio, lo que ocurriría en el caso de los delirios.

Críticas al enfoque de Campbell

La propuesta de Campbell ha sido objeto de numerosas críticas. Sintetizaré a continuación aquellas que, a mi criterio, muestran que el enfoque *"bottom-up"* se encuentra en mejores condiciones para proveer un modelo explicativo sobre los delirios. Como puede apreciarse, el enfoque de Campbell implica dos afirmaciones principales. La primera es que existe algún disturbio de tipo "*top-down"* en algunas creencias fundamentales del sujeto, lo cual afectaría su experiencia perceptual[3]. La segunda es la idea de que los delirios tendrían el mismo status que las proposiciones marco de Wittgenstein. Ambas afirmaciones son lógicamente independientes. Una creencia podría funcionar como una proposición marco sin mantener un vínculo "*top-down"* con la experiencia y, a su vez, una creencia podría

mantener un vínculo "*top-down"* sin ser una proposición marco.

Detengámonos en la primera afirmación que refiere a la etiología del delirio. Según vimos, Campbell (2001) considera que las creencias delirantes emergen como resultado directo de un mal funcionamiento orgánico[4]. Sin embargo, como han señalado Bayne y Pacherie (2004), esto presenta una desventaja (aunque no es una objeción de principio) respecto al enfoque "*bottom-up*" pues es comúnmente aceptado, como el mismo Campbell reconoce, que un cambio en la experiencia perceptual podría explicarse como un resultado directo de un mal funcionamiento orgánico, pero no parecer haber aval similar para la idea de que un mal funcionamiento orgánico podría alterar determinadas creencias específicas, como ocurre en los casos de los delirios monotemáticos. En segundo término, Campbell (2001) critica al enfoque afirmando que no es sensato considerar a la creencia delirante como una "respuesta racional" a la experiencia afectiva anómala[5]. Ahora bien, aunque es válido señalar que bajo el enfoque "*bottom-up*" podría no resultar del todo claro cómo el sujeto pasa desde una experiencia afectiva anómala a un contenido perceptual específico y, desde allí hacia una creencia delirante, esta crítica se aplica solo a los modelos de un factor. Como veremos luego, un modelo multifactorial puede ser capaz, añadiendo otros factores, de explicar (siquiera parcialmente) por qué el sujeto llega a la creencia delirante.

Por otra parte, un problema adicional para Campbell es explicar por qué los delirios monotemáticos estarían circunscriptos a temas particulares, como de hecho lo están. Si las creencias delirantes emergen directamente de cierto daño cerebral ¿por qué, por ejemplo, no nos encontramos con muchos tópicos delirantes diferentes? ¿Por lo que se altera son únicamente los significados de creencias con contenidos específicos? Un enfoque de tipo "*top-down"* no parece contar con los recursos para dar cuenta de la evidencia que muestra que los delirios monotemáticos tales como el Capgras o el Cotard están restringidos a tópicos muy específicos.

Pasemos ahora a considerar la segunda afirmación de Campbell. El autor sostiene que las creencias delirantes tendrían el mismo status que se les asigna a las denominadas proposiciones marco. Ahora bien, aunque el propio Campbell no lo pone exactamente en estos términos, es preciso señalar que, de acuerdo a Wittgenstein (1969), estas proposiciones parecer tener características distintivas tanto de tipo epistémico como semántico.

Ahora bien, si estos son los rasgos distintivos de las proposiciones marco, es posible analizar la sugerencia de Campbell, focalizándonos por un lado en si los delirios son realmente inmunes al escrutinio empírico ordinario y, por el otro, si el significado de las palabras de los pacientes de hecho ha cambiado. Respecto al primer punto, cabe señalar, inicialmente, que no es cierto que los delirios no respondan en absoluto a la evidencia pues, por un lado, los pacientes en algunas ocasiones consideran cierta evidencia que podría interpretarse a favor de sus delirios. Como afirma el paciente Schreber (1979): "Existe un cúmulo abrumador de razones probatorias, que desearía que usted conociese en detalle por el contenido de mis *Memorias*" (p. 10). Por otro lado, si tomamos en cuenta los resultados de las terapias cognitivo-conductuales (TCC), la evidencia sugiere que pueden ser efectivas al menos para el tratamiento de algunos delirios (Coltheart, 2007; McKay y Cipolotti, 2007; Nelson, 2005). Este enfoque, basado en la idea de que los delirios son un tipo de creencias propone como una de las técnicas principales cuestionar las creencias delirantes del paciente a la luz de contra-evidencia, de modo progresivo y gradual (Dickerson, 2000). La (TCC) ha proveído efectivo tratamiento de algunos delirios y tal efectividad muestra que los delirios responden, al menos algunas veces, a la evidencia (Garety y Hemsley, 1997; Read, Mosher y Bentall, 2005). En el proceso terapéutico, en el cual se discute aquello que puede interpretarse como evidencia y se modifique su peso, mostraría que los sujetos delirantes no necesariamente se mantienen con una convicción inquebrantable, sino que esta puede variar gradualmente. Asimismo, cabe señalar que los pacientes delirantes son a menudo capaces de apreciar la implausibilidad de sus creencias y en este sentido se mostrarían predispuestos a intentar someterlas a escrutinio, al menos progresivamente (Bayne y Pacherie, 2004; Young, 1998).

En cuanto al aspecto semántico, ¿ha cambiado realmente el significado de las palabras que utilizan los pacientes cuando describen o expresan sus delirios? ¿Cómo sería posible que el paciente pudiera cambiar los significados de algunas palabras solo en el contexto de la formulación de sus creencias delirantes? ¿Es realmente plausible que el significado que da un sujeto a sus palabras sea sensible al contexto de este modo? Si bien el significado de nuestras palabras a veces puede ser sensible al contexto (lingüístico y no lingüístico), pensar que el significado de las palabras

pueda alterarse solo cuando el sujeto expresa su delirio, no resulta demasiado plausible. Como bien señalan Bayne y Pacherie (2004, p. 7) respecto a un paciente con Capgras, si el paciente dice "esta mujer no es mi esposa", ¿qué otra cosa podría querer decir si no que la mujer que está mirando no es la misma persona que la mujer que él tiene en sus memorias? Acuerdo aquí con los autores en que no es claro que haya una mejor caracterización disponible del significado de los términos empleados por el sujeto delirante. Asimismo, es difícil reconciliar la tesis de que los sujetos delirantes han cambiado el significado de los términos que usan con el hecho de que muchos pacientes son capaces de comprender que otros pueden encontrar difícil de creer su propia historia. Finalmente, cabe preguntarse, ¿pueden efectivamente cambiarse las proposiciones marco? Klee (2004) y Thorton (2008) han criticado este aspecto de la propuesta de Campbell señalando que esto sería implausible bajo las ideas del propio Wittgenstein, pues pareciera que cada hablante de un lenguaje podría escoger cuáles proposiciones marco constituyen su marco cognitivo personal. Asimismo, esto implicaría el error de pensar en términos de "significados privados", algo que va en contra de la idea wittgensteiniana de la naturaleza pública del significado. Como menciona Klee (2004) el sujeto delirante no cambia (no puede cambiar) el significado de un término de un lenguaje público meramente por un acto individual de voluntad imaginativa que involucre hacer aserciones bizarras. La naturaleza pública del significado es clave en la visión wittgensteiniana, de modo que significar algo a través del uso de palabras en un lenguaje público no puede ser equivalente a un acto de imaginación individual, pues el significado de un término es fijado por su uso público actual.

En síntesis, las críticas mencionadas hacia el enfoque "*top-down*" nos brindan razones para pensar que dicho enfoque no brinda una explicación adecuada de los delirios. Esto se debe a que, en primer lugar, no resulta claro cómo un mal funcionamiento orgánico podría generar cambios en las creencias específicas del sujeto o cómo podrían constituirse los contenidos específicos del delirio. En segundo lugar, la comparación con las proposiciones marco tampoco parece ser iluminadora pues no resulta claro que los delirios tengan el mismo estatus semántico y epistémico que dichas proposiciones wittgensteinianas.

El enfoque *"bottom-up"*

Los enfoques de tipo "*bottom-up"*, han ganado prioridad en la literatura por resultar más ajustados empíricamente (Davies y Coltheart, 2000). Para los defensores de este enfoque, los delirios implican modificaciones del sistema de creencias que son producidas por experiencias perceptuales inusuales y extrañas, generadas por un mal funcionamiento de tipo orgánico (Bayne y Pacherie, 2004; Davies *et al.*, 2002; Ellis y Young, 1990; Maher, 1974; Maher, 2001; McKay, Langdon y Coltheart, 2009; Stone y Young, 1997). Este enfoque implica tres tesis centrales: (1) que la causa proximal de la creencia delirante es cierto tipo de experiencia inusual, (2) que la creencia delirante es una respuesta racional a la experiencia inusual del paciente y (3) los términos que el sujeto emplea para expresar sus creencias preservan el significado usual (Bayne y Pacherie, 2004).

Los defensores de este enfoque, se dividen en dos grupos principales: aquellos que consideran que la experiencia perceptiva inusual es suficiente para la formación del delirio (la teoría de un factor) (Maher 1974, 2001) vs. aquellos que consideran que la experiencia perceptiva inusual es solo un factor en la formación del delirio, pero postulan otro factor adicional (la teoría de los dos factores) (Davies *et al.*, 2002; Ellis y Young, 1990; Stone y Young, 1997).

Respecto al enfoque de un factor, Maher (1974, 1988, 1992, 1999, 2001) ha ofrecido un modelo de formación de delirios que enfatiza el rol causal de la experiencia anómala y afirma que esta sería suficiente para generar y mantener el delirio. A criterio de Maher (2001), el pensamiento delirante no es en sí mismo cognitivamente aberrante, es decir que los procesos cognitivos por los cuales se forman los delirios no son diferentes, en un modo relevante, de aquellos que forman las creencias normales. Por otra parte, se afirma que los delirios serían como las teorías científicas en la medida en que sirven al propósito de proveer orden y significado a las observaciones empíricas y, como en el caso de la teorización científica, la necesidad de una teoría emerge cuando la naturaleza se nos presenta como desconcertante. Es decir, cuando no tienen lugar los eventos que predijimos, cuando aparecen eventos impredecibles y la observación discrepa con nuestras expectativas.

Puesto que los delirios no serían significativamente diferentes de los

procesos de formación de creencias normales, Maher (2001) sostiene que, si los delirios son patologías de las creencias, entonces el *locus* de la patología descansa en la experiencia y no en los procesos de razonamiento del sujeto. Cuando la experiencia anómala ocurre, esta atrae la atención del sujeto y se produce tensión. Esta tensión, motivaría la búsqueda de una explicación, que no se detiene hasta ser hallada. Si bien la explicación puede no ser completamente adecuada, reducirá la ansiedad y generará alivio, lo que hace que sea mejor una explicación defectuosa o incompleta antes que ninguna explicación. Desde este marco, los delirios serían respuestas normales y el sujeto delirante es visto como ampliamente racional en lo relativo al proceso de la formación del delirio.

En la literatura se han señalado diversas críticas a la teoría de Maher (Davies *et al.*, 2002; Porcher, 2015). Por un lado, encontramos una crítica a la insuficiencia de esta explicación para dar cuenta de la formación del delirio; por otro lado, una crítica similar que refiere a la insuficiencia de la explicación para dar cuenta del hecho de que el delirio se mantenga y no sea descartado frente a evidencia contraria.

El primer tipo de críticas, señala que las experiencias anormales no siempre generan un delirio y que, además, muchos delirios ocurren en ausencia de cualquier experiencia anómala (Davies *et al.*, 2002; Porcher, 2015). Al respecto, una revisión de estudios realizada por Davies y colegas (2002) aporta evidencia que sugiere, en contra de Maher, que un solo factor no puede ser suficiente y que, por lo tanto, habría otros factores involucrados en la formación del delirio. De acuerdo al enfoque de un factor se seguiría que cualquiera que sufre de un daño neuropsicológico que por ejemplo reduzca su respuesta afectiva a los rostros, debería exhibir un delirio de Capgras. Sin embargo, estas predicciones han sido falsadas por ejemplos que provienen de la literatura neuropsicológica que muestran que los sujetos, a pesar de tener una experiencia anómala, no forman explicaciones delirantes de su experiencia (Davies *et al.*, 2002; Porcher, 2015).

El segundo tipo de críticas apunta a señalar la insuficiencia de la experiencia anómala para explicar el mantenimiento del delirio, un aspecto que cualquier modelo adecuado debería explicar. Esto se debe a que, incluso si una experiencia anómala provee una respuesta a la cuestión sobre cuál es el origen del delirio, no provee ninguna respuesta a la cuestión de

por qué el paciente no rechaza la creencia delirante ante la evidencia en contra de la misma. En consecuencia, también es preciso dar cuenta del porqué de se mantiene el delirio y, para ello, debe postularse un segundo factor.

Al respecto, se han desarrollado diversos modelos (Davies *et al.*, 2002; Ellis y Young, 1990; Stone y Young, 1997), los cuales, aunque coinciden en que es necesario postular un segundo factor, difieren en cómo este debe ser caracterizado. Pese a no haber acuerdo respecto a este punto, Davies *et al.*, (2002) y colegas señalan que el segundo factor podría entenderse de modo general en términos de "la incapacidad para rechazar un candidato a una creencia sobre la base de su implausibilidad e inconsistencia (aunque sea parcial) con el resto de lo que el paciente conoce" (p. 149). Como candidatos a segundo factor, se han postulado déficits y factores motivacionales.

El segundo factor: motivaciones

Bentall y colegas han propuesto que los delirios (en particular los persecutorios) serían el resultado de sesgos en el tipo de explicaciones que los sujetos dan de su propia conducta y de la conducta de otras personas, pues los pacientes tenderían culpar a las otras personas de las cosas negativas que les suceden (Bentall, 1995; Bentall y Kinderman, 1998). En la misma dirección, Stone y Young (1997) han propuesto una explicación de dos factores, para los delirios de Capgras y Cotard, en la cual el segundo factor consiste también en un sesgo atribucional. En este modelo, la diferencia fundamental entre los pacientes tiene que ver con el estilo atribucional que el individuo es propenso a adoptar. Si predomina un sesgo externalizante, entonces el sujeto atribuye la culpa de sentir ese sentimiento de no-familiaridad a factores externos, por lo tanto, adopta la explicación del Capgras de acuerdo a la cual el cónyuge ha sido reemplazado por un impostor. Si en cambio adopta un sesgo internalizante, entonces el sujeto atribuye la culpa a factores internos, adoptando la explicación del Cotard, de acuerdo a la cual el sujeto está muerto o en algún sentido es irreal. En consonancia, Young (2000) ha sugerido lo mismo, señalando que, aunque estos delirios son fenomenológicamente diferentes, pueden representar

distintos intentos de los pacientes de dar sentido a experiencias similares.

Así, los sesgos atribucionales podrían explicar, en algunos casos, por qué una hipótesis delirante particular es priorizada como un candidato explicativo de una experiencia inusual. Por otra parte, otros autores han apelado a sesgos de razonamiento, como un tipo de sesgo más general que podría incluir distintos subtipos. Al respecto, Garety y colegas sostienen que un factor en la etiología de los delirios es la tendencia a saltar precipitadamente a las conclusiones pese a tener evidencia insuficiente para ello (“*jumping to conclusions bias”*) (Garety, Hemsley y Wessely, 1991; Garety y Freeman, 1999; Dudley, John y Young, 1997). En este sentido, aunque los resultados en los experimentos de razonamiento probabilístico con los sujetos delirantes no revelan una completa inhabilidad para razonar probabilísticamente, sí muestran un sesgo hacia la aceptación precipitada de hipótesis. En consonancia, Coltheart, Langdom y McKay (2011) han argumentado a favor de un modelo de dos factores, señalando que ningún estilo de razonamiento puede ser tan sesgado como el de los sujetos delirantes y que, consecuentemente, la etiología de los delirios debe incluir un déficit en el razonamiento, que ha de sumarse a las experiencias anormales y a los típicos sesgos de razonamiento. Así, a criterio de los autores, los delirios emergerían cuando el sistema cognitivo normal que las personas usan para generar, evaluar y adoptar creencias, se encuentra dañado. No obstante, aún no resulta claro cuáles serían los déficits o sesgos específicos de razonamiento, ni qué nivel de déficit es necesario para que un delirio persista. Como han señalado Davies *et al.* (2002), la evidencia sugiere que el daño de las capacidades de revisión de creencias del sujeto delirante es parcial, ya que incluso en la mayoría de los casos psicóticos más dramáticos, hay aún alguna preservación del razonamiento normal. Luego, sería necesario seguir profundizando en hipótesis e investigaciones empíricas en pos de poder caracterizar más adecuadamente los sesgos en el razonamiento, determinar de qué tipos de sesgos se trata y en qué grado deben estar presentes para que el delirio se mantenga en el tiempo.

Por su parte, los enfoques motivacionales consideran que la producción del delirio está causada por los beneficios psicológicos que este conferiría al sujeto (Bentall y Kaney, 1996; Bell, 2003; Bortolotti, 2014). Desde esta visión, el fenómeno delirante es caracterizado como una respuesta psicológica activa a una amenaza interna o a un estímulo psicológico

externo. En la literatura actual, los enfoques motivacionales ven a los delirios como un modo de lidiar con ciertas situaciones conflictivas que resultan abrumadoras para el sujeto (Bentall, Corcoran, Howard, Blackwood y Kinderman, 2001; Pavlickova *et al.*, 2013). Al respecto, se ha señalado que los delirios psicológicamente motivados ayudarían a manejar emociones negativas fuertes y a incorporar las experiencias negativas a la vida psíquica del sujeto (Bentall, 1994; 2003; McKay, Langdom y Coltheart, 2009). Por ejemplo, en el caso del Capgras los factores emocionales podrían estar asociados a la naturaleza selectiva del delirio y contribuirían a la distorsión de los datos perceptivos, provocando falsas creencias sobre las personas más relevantes para este. En este sentido, resulta importante no considerar a los enfoques del déficit y a los motivacionales como excluyentes, sino avanzar en el desarrollo de modelos que combinen ambos tipos de factores en las explicaciones sobre la génesis y mantenimiento de los delirios.

El proceso psicopatológico del delirio

Pues bien, aunque los modelos de tipo "*bottom-up*" resultan empíricamente más ajustados y han conducido al desarrollo de distintos enfoques de dos factores que consideran, además de la experiencia anómala, ciertos sesgos, déficits de razonamiento o factores motivacionales en la producción del delirio, hablar solo de dos factores parece resultar insuficiente para dar cuenta de la complejidad de la producción y mantenimiento del delirio, especialmente porque esto deja de lado componentes importantes en el proceso. Por un lado, se deja de lado –al menos parcialmente– el importante rol de los afectos en la génesis y el mantenimiento de los delirios. Por otra parte, se pretende explicar solo la creencia delirante dejando de lado el contexto psicopatológico en la que esta emerge y la relación que mantiene con la vida psíquica global del sujeto.

Lejos de esto, si se pretende comprender adecuadamente al delirio, como bien han señalado los primeros psicopatólogos (Ey, 1950), no hay que considerar a la creencia delirante despojada del contexto psicopatológico en el que emerge, analizándola como un fenómeno aislado y poniendo énfasis solo en las características epistémicas del fenómeno. Antes bien, debe

considerarse el carácter procesual del fenómeno psicopatológico, en este caso el delirio que abarca desde experiencias primarias agudas, de carácter perceptual-afectivo, hasta el momento en el cual se torna crónico, donde las creencias delirantes se cristalizan e integran a la vida psicológica del sujeto. Esto quiere decir además que el delirio es un aspecto, un síntoma, de un trastorno psicopatológico mayor que lo engendra (que podrá ser uno o diversos). O, para utilizar una simple metáfora, el delirio puede ser visto como la punta de un enorme iceberg. En tanto reconozcamos el carácter procesual, comprenderemos que nos referimos a un conjunto de diversas etapas sucesivas, en donde podemos identificar, antes del *breakdown* cognitivo, una fase prodrómica[6], en la cual encontramos ciertas características llamativas (las experiencias agudas) que preceden a la aparición de las creencias delirantes. En esta fase, a lo que deberíamos atender es a un aspecto que, si bien ha sido considerado en la tradición psicopatológica, se ha descuidado en los recientes debates filosóficos sobre los delirios. Me refiero a ciertos aspectos afectivos de carácter fenomenológico que se producen en la atmósfera previa a la generación del delirio (Freeman *et al.*, 2002; Freeman, 2016).

El trasfondo de afectividad negativa y la atmósfera pre-delirante

Diversos estudios reportan que los delirios emergen en un contexto general de afectividad negativa (Kramer *et al.*, 2013; Strauss *et al.*, 2013). Sin embargo, este aspecto ha sido descuidado en los recientes debates filosóficos sobre los delirios, los cuales suelen centrarse solo en las características epistémicas del fenómeno. En general, puede decirse que la comprensión de los trastornos afectivos por parte de la psiquiatría actual es aún escasa, debido al insuficiente entendimiento de la naturaleza de la afectividad (Berrios, 2008).

La evidencia parece indicarnos que los diferentes dimensiones de la afectividad se encuentran dañadas en el período que precede a la emergencia del delirio (Strauss *et al.*, 2013; Fuchs, 2013; Marwaha, Broome, Bebbington, Kuipers y Freeman, 2013; Maiese, 2013). Se ha encontrado que los pacientes que desarrollan síntomas psicóticos usualmente muestran un humor pre-psicótico y alta inestabilidad afectiva

(Strauss *et al.*, 2013). Marwaha y colegas (Marwaha *et al.*, 2013) han identificado un vínculo robusto entre la inestabilidad afectiva y los síntomas psicóticos, concluyendo que la primera puede predecir la emergencia de ciertos delirios. En la misma línea, se han reportado altas tasas de ansiedad y estados depresivos entre personas con riesgo de psicosis, con anterioridad a la ocurrencia de algunos estados psicóticos tales como los delirios y las alucinaciones auditivas (Bentall *et al.*, 2001; Broome, Woolley y Tabraham, 2005; Broome, Day yValli, 2012; Freeman *et al.*, 2002). Estas emociones negativas juegan un rol importante en la génesis de los delirios, y diversos estudios han encontrado una elevada presencia de las mismas en individuos con síntomas positivos de psicosis, tales como delirios (para una revisión véase Turnbull y Bebbington, 2001). Adicionalmente, muchos de los procesos implicados en el mantenimiento de trastornos de ansiedad (Clark, 1999) parecen estar implicados también en el mantenimiento de los delirios persecutorios (Freeman y Garety, 1999; 2002). Asimismo, se ha señalado que sentir que las experiencias emocionales están fuera del control, puede impulsar a la búsqueda de significados que puedan encontrar explicaciones en términos de influencias externas (Marwaha *et al.*, 2013). En síntesis, podemos decir que hay evidencia para pensar que los delirios emergen en el contexto de un trasfondo de afectividad negativa (O'Driscoll, Laing y Mason, 2014).

Ahora bien, las etapas tempranas de las psicosis han atraído la atención de los psicopatólogos desde los comienzos de la psiquiatría (Bleuler, 1911; Ey, 1950) pues se dice que en estas el ambiente es percibido por los pacientes como desconcertante, misterioso y como si todo fuera similar a una "puesta en escena" y que los objetos, o ciertas personas, pueden aparecer como teniendo una expresión abrumadora (Fuchs, 2005).

Como ha señalado Jaspers en su *magnum opus*, *Psicopatología general* (1913/1963), aparecen en los enfermos sensaciones primarias, sentimientos vivaces, disposiciones de ánimo y cogniciones: "Pasa algo, dime qué es lo que pasa", así se expresó una paciente de Sandberg ante su esposo. Al preguntarle qué ocurría, sostenía la paciente: "Sí, no lo sé, pero hay algo". Para los enfermos es de mal agüero que ocurra algo que ellos presienten, pues todo tiene para ellos una nueva significación. El ambiente es distinto, "existe una alteración que se acerca, sutil, que lo envuelve todo con una luz incierta, de mal aspecto. Una habitación antes indiferente o amable provoca

ahora un estado de ánimo indefinible" (p. 122).

El periodo previo a la adopción del delirio es caracterizado por ciertos cambios perceptuales y afectivos que generan una transformación fenomenológica general del modo en el cual los pacientes perciben y sienten el mundo (Jaspers 1963, p. 97). Para comenzar, la atmósfera experimentada en estas etapas iniciales es caracterizada por una alienación radical del ambiente percibido y se experimentan alteraciones afectivas que "colorean" la experiencia entera del mundo con un carácter comúnmente abrumador y negativo (Gibbs, 2000; Fuchs, 2013).

Como señala Fuchs (2005) estas atmósferas pueden ser consideradas como cualidades afectivas holísticas de la experiencia espacial y de las situaciones interpersonales, que integran sus características expresivas en una *gestalt* dinámica unitaria: por ejemplo, el sentimiento de alegría en una fiesta, la tristeza de un funeral, el clima frío de una conferencia, etc. Como todos los fenómenos afectivos, las atmósferas son experimentadas a través de una resonancia en el cuerpo. Así, por ejemplo, un clima de tensión interpersonal se siente como opresivo o sofocante (Fuchs 2005, p. 616). En comparación con los sentimientos existenciales o los humores, las atmósferas afectivas a menudo se sienten como un aura envolvente, extraña, artificial y desconcertante que es emanada desde el ambiente en el que uno se encuentra (Anderson, 2009). Es así que todo parece arreglado, como si se tratase de un escenario, y las cosas dan la impresión de ser meras imitaciones con un propósito que no es posible descifrar. Los pacientes a menudo describen experiencias como las siguientes*:* "Todo es extraño, o todo es en algún aspecto diferente" (Saas, 1992, p. 46), "cuando vas a algún lado, todo parece ya preparado para ti como en un teatro, es realmente misterioso y te asustas terriblemente" (Fuchs, 2005, p. 134). Es así que se genera un sentimiento general de incertidumbre que no solo impregna la experiencia global de realidad, sino también los objetos específicos en el campo de la conciencia (Freedman y Chapman, 1973). Como consecuencia, los pacientes comienzan a experimentar una tensión opresiva que se va incrementando, sentimientos de incertidumbre, desunión y falta de control sobre sus experiencias del mundo y sobre su propia vida mental (Mishara, 2010).

En el temple delirante, como lo denomina Jaspers (1963) *"Wahnstimmung"*), hay siempre algo que es germen de un valor y

significación objetivas. Los pacientes sufren y, en consecuencia, la adquisición de una explicación determinada resulta un alivio ya que la repentina conciencia de una idea clara, sea verdadera o falsa, tiene *per se* ya un efecto tranquilizador. Es decir, que en medio de este extrañamiento ocurre un segundo cambio: los objetos o las personas singulares ganan un nuevo, misterioso y desconcertante significado (Fuchs, 2005, p. 134) que logra aliviar al menos parte de la tensión y la angustia que le generaban las experiencias desconcertantes.

Pese a ser un fenómeno ubicuo, la atmósfera de la afectividad ha sido raramente tratada en las investigaciones, no solo psicopatológicas, sino incluso en aquellas sobre las emociones ordinarias (Fuchs, 2005). Resulta importante, sin embargo, incluir a la atmósfera delirante en los modelos que se pretendan desarrollar sobre la generación de los delirios. Respecto a la valoración clínica, puede decirse que si el estado pre-delirante va a jugar en el futuro un papel en el diagnóstico y tratamiento, o a penetrar en el conocimiento de los mecanismos, es fundamental que el fenómeno clínico sea capturado. Como señalan Berríos y Fuentenebro (1995), la pérdida de interés hasta ahora percibida por la psiquiatría hacia el estado pre-delirante ha estado en relación con su situación refractaria al análisis directo (p. 148). La comprensión y evaluación de la atmósfera pre-delirante requiere de la creación de un lenguaje que pueda identificar las cualidades y continuidades de la experiencia en sí misma, su tono emocional, duración, etc., así como que el paciente sea examinado lo más pronto posible pues es probable que las evaluaciones retrospectivas sean menos informativas, pese a lo cual, en muchas ocasiones, constituyen la única información disponible. Por todo ello, es necesario profundizar el estudio de los estados pre-delirantes y el peso que tienen determinados afectos negativos en la generación y mantenimiento del delirio. Pasemos ahora a considerar cómo todos los factores mencionados hasta aquí podrían integrarse en un modelo multifactorial de tipo "*bottom-up*", basado en el carácter explicativo que tenga en consideración la etapa pre-delirante y los factores afectivos en la emergencia del delirio.

Un modelo multifactorial

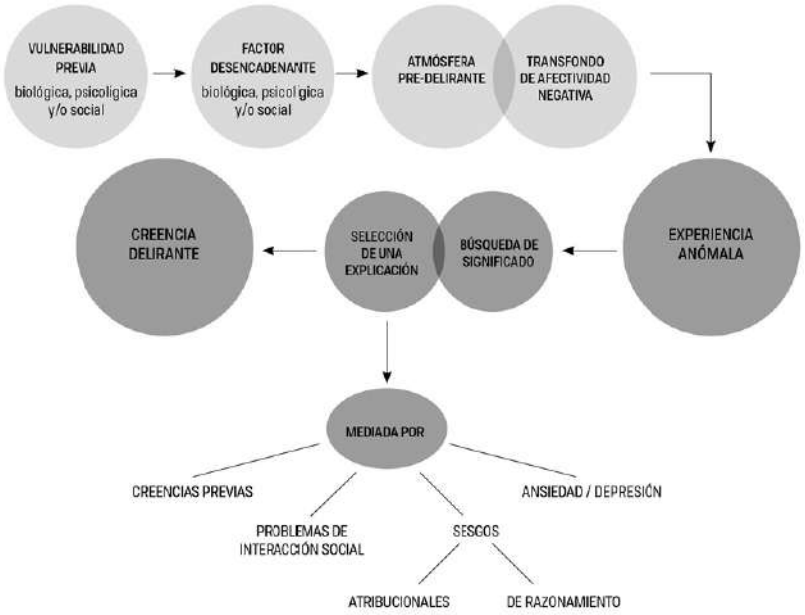
Si hablar de solo dos factores parece resultar insuficiente para dar cuenta de la complejidad de la producción y mantenimiento del delirio, entonces parece necesario desarrollar modelos multifactoriales para dar cuenta de la complejidad del fenómeno, que resulten teórica y clínicamente más fructíferos. En esta línea, Garety y colegas (Garety *et al.*, 2001) han propuesto un modelo multifactorial de carácter general para los síntomas positivos presentes en las psicosis, el cual integra los niveles psicológicos, neurobiológicos y sociales para generar hipótesis sobre los procesos causales de las psicosis. Más allá de las ventajas del modelo, y de que los síntomas positivos de la psicosis frecuentemente co-ocurren como bien señalan Freeman *et al.* (2002), es necesario desarrollar modelos específicos para los distintos tipos de síntomas, en nuestro caso, el delirio.

Los modelos de múltiples factores en psicopatología suelen apoyarse en un marco conceptual general de vulnerabilidad-estrés (también denominado modelo "diátesis-estress"). Este es un marco heurístico sobre el origen de las psicosis que intenta ofrecer una integración teórica de aquellos factores que conforman el desarrollo y curso de las psicosis y sus síntomas principales. Este marco conceptual ha sido originalmente propuesto por Zubin y Spring (1997) y, desde entonces, se ha avanzado en el desarrollo de distintos modelos para diversos trastornos[7].

Desde dicho marco se parte del supuesto de que existe una predisposición a las psicosis de naturaleza genética o adquirida (por ejemplo, como consecuencia de daño cerebral en el desarrollo). La vulnerabilidad, sin embargo, no se considera suficiente para manifestar el trastorno, sino que necesita ser desencadenada por procesos ambientales. Los componentes ambientales pueden ser biológicos (p. ej.: una infección), psicológicos (p. ej.: situaciones vitales estresantes) y/o sociales (p. ej.: pérdida del empleo, discriminación, etc.) La intensidad del estrés ambiental necesario para desencadenar el proceso difiere de una persona a otra, y lo mismo ocurre con el grado de vulnerabilidad. Asimismo, existen factores de protección tanto personales (p. ej.: la inteligencia, las habilidades sociales o la capacidad de resiliencia) como ambientales (ej: el apoyo familiar o social), los cuales pueden atenuar el riesgo de psicosis en personas constitucionalmente predispuestas. Por otra parte, se reconoce que habría factores estresantes que incidirían en el desencadenamiento del delirio, tales como sucesos vitales estresantes (ej: duelos), consumo de drogas o

condiciones de vida problemáticas (p. ej.: pobreza, conflictos familiares, abandono, etc.) (Goh y Agius 2010; López y Vázquez, 1998).

Los modelos para los síntomas específicos, en particular para los delirios, pueden ser encuadrados en este marco conceptual general de vulnerabilidad-estrés, donde se asume que la emergencia de los síntomas depende de la interacción entre cierta vulnerabilidad (desde factores biológicos, psicológicos y sociales) y el estrés (el cual también puede ser biológico, psicológico y social). Un modelo multifactorial para los delirios puede ser enmarcado en este marco conceptual de vulnerabilidad-estrés, en el cual se asume que la emergencia del delirio depende de la interacción entre cierta vulnerabilidad previa y el estrés de la situación actual del sujeto. Así, la formación del delirio comenzará a partir de un desencadenante, tal como podría serlo un evento vital estresante. A partir de allí, comenzaría la etapa prodrómica, en la cual aparece lo que se ha denominado la atmosfera delirante, la cual se encuentra acompañada de un trasfondo de afectividad negativa (principalmente ansiedad y depresión mantenidas durante un largo período de tiempo) (Fowler, 2000), tal como la hemos descripto previamente. Eventualmente, se termina desencadenando una experiencia anómala. Esta puede consistir en experimentar voces y/o acciones de uno como ajenas, ciertas anomalías perceptuales, pérdida de sentimientos de familiaridad, sensaciones corporales anómalas, etc. (Freeman, Garety y Philips, 2000). Todas estas experiencias, como mencionamos, generan en el paciente un estado de tensión y ansiedad que terminarán conduciéndolo a la búsqueda de significado, a la búsqueda de una explicación que lo tranquilice y alivie la tensión.



En esta búsqueda de significado, se ha señalado que las diversas creencias pre-existentes (sobre sí mismo, sobre los otros y sobre el mundo) influenciarían la construcción de la explicación delirante (Freeman, 2016; Trower y Chadwick, 1995). Por otra parte, la evidencia sugiere que este proceso de búsqueda de significado se ve influenciado, además, por distintos tipos de sesgos cognitivos: sesgos generales de razonamiento[8](se destacan el sesgo de salto a las conclusiones, la necesidad de cierre que hace que los sujetos no toleren la ambigüedad de los estímulos y el sesgo de confirmación, que hace que los sujetos busquen información que tienda a confirmar sus creencias) (Garety *et al.*, 1991; Garety y Freeman, 1999; Dudley *et al.*, 1997; Coltheart, Langdom y McKay, 2011) y sesgos atribucionales (Bentall, 1995; Kinderman, 1994; Bentall y Kinderman, 1998).

Asimismo, se ha señalado que este proceso podría estar influenciado por problemas relacionados a las interacciones sociales, ya sean ciertas disfunciones en teoría de la mente (Firth, 1992; Langdon, Davies y Coltheart, 2000) o algún tipo de problemática psicológica-social que hace que los sujetos no confíen en otros agentes cuya opinión debería ser de fiar (Davies *et al.*, 2002). Se ha sugerido, por ejemplo, que, si la persona está aislada, puede volverse incapaz de revisar sus pensamientos sobre la base de la interacción con otros significativos que lo apoyen, con lo cual los delirios de amenaza, por ejemplo, florecen más fácilmente (Freeman *et al.*,

2002). Adicionalmente, se ha conjeturado que los niveles de ansiedad y depresión pre-existentes también influenciarían qué tipo de explicación resulta seleccionada. Por ejemplo, una baja auto-estima, junto a niveles elevados de ansiedad, podría conducir a ciertos delirios persecutorios (Mathews, 1990; Freeman y Garety, 2000).

En síntesis, podemos decir que los delirios emergen como resultado de la búsqueda de una explicación a una experiencia anómala, reflejando esto una interacción entre factores de distintos órdenes (los asociados a la vulnerabilidad previa del sujeto, los factores estresantes desencadenantes, los factores afectivo-emotivos, los sesgos cognitivos y los problemas de interacción social), lo cual apunta a explicaciones de carácter híbrido y pluralistas. La evidencia sugiere que todos estos factores no son independientes, no obstante, deben llevarse a cabo investigaciones empíricas que nos permitan estudiar los vínculos y las múltiples influencias que existen entre los mismos. A continuación, se presenta el esquema del modelo propuesto:

Ahora bien, un buen modelo debería servir, siquiera heurísticamente, como guía para comprender también por qué se mantiene el delirio. Al respecto, considero que hay una respuesta de carácter general que podría ayudar a responder a esta pregunta para los distintos tipos de delirios. Sin embargo, al tener un carácter general, posiblemente sirva como una guía heurística y deberá testearse y analizarse empíricamente para especificar lo que sucede en los distintos subtipos de delirios. A partir de la evidencia disponible, esta respuesta de carácter general apunta a considerar múltiples factores cognitivos que colaboran con el mantenimiento del delirio. Se han hallado distintos tipos de sesgos de razonamiento y atribucionales como los ya mencionados. Se ha registrado la influencia de ciertos estilos cognitivos (en particular el estilo *auto-referencial)* que afectarían el proceso (Freeman, Garety y Phillips, 2000). Así mismo, habría sesgos que llevan a descartar la evidencia disconfirmatoria (Moritz y Woodward, 2006; Woodward, Moritz y Chen, 2006), además, diversos sesgos atencionales (Bentall y Kaney, 1996) y sesgos de memoria (Freeman *et al.*, 2002) y problemas en el uso de los conocimientos acumulados (Davies *et al.*, 2002).

En síntesis, diversos estudios empíricos avalan la idea de que distintos sesgos cognitivos se encuentran involucrados en el mantenimiento de los delirios (Moritz y Woodward, 2005; Moritz, Woodward y Lambert, 2007;

Moritz y Woodward, 2006; Woodward *et al.*, 2008; Veckenstedt *et al.*, 2011; Lincoln, Schäfer y Moritz, 2010). Esos mismos estudios apoyan la idea de que se trata de algo que difiere en *grados* respecto de los sujetos del grupo control (es decir el grupo de sujetos sin trastornos psicopatológicos), y no cualitativamente (Moritz y Woodward, 2005).

Por otra parte, en relación a los factores motivaciones, se ha destacado el componente emocional, señalando que habría ciertas "*conductas de seguridad*" por parte de los sujetos delirantes que los previenen de un potencial daño psicológico. Dichas conductas serían estrategias de afrontamiento que desarrollan los sujetos para reducir el malestar psicológico frente a ciertos estímulos y/o situaciones. En el caso de los sujetos delirantes, un ejemplo de ellas consistiría en dar respuestas hostiles a la gente que contradice sus delirios (Milosevic y Radomsky, 2008; Salkovski, 1996).

Así también se han destacado ciertas alteraciones sociales, señalándose, en particular, que las interacciones con los demás se ven alteradas, ya sea por aislamiento u hostilidad (Langdon *et al.*, 2002; Freeman *et al.*, 2002; Wahl, 1999). A su vez, Davies y colegas (Davies *et al.*, 2002) también han señalado que habría posibles déficits en la teoría de la mente o la cognición social (p. 151) que previenen la obtención de evidencia disconfirmatoria.

Observaciones finales

En este escrito me he dedicado a examinar diversos puntos del debate en torno al problema de cómo desarrollar un modelo explicativo adecuado para la formación y el mantenimiento de los delirios. Por una parte, he presentado los dos enfoques explicativos de los delirios dominantes: el enfoque "*top-down*" y el enfoque "*bottom-up*". Por otra, he considerado los diversos factores que estarían involucrados en un modelo explicativo sobre los delirios. Al respecto, he defendido la necesidad de desarrollar modelos de tipo "*bottom-up*", basados en el carácter explicativo y de carácter multifactorial. De tipo "*bottom-up*", en la medida en que la evidencia empírica parece apoyar la idea de que la dirección la causalidad en la génesis del delirio es de abajo hacia arriba, es decir, que va desde la experiencia anómala hacia la formación de la creencia delirante. Es un

modelo basado en el carácter explicativo pues se considera que se adopta el delirio en un intento por explicar una experiencia altamente inusual. Por último, se trata de un modelo de tipo multifactorial que incorpora diversos factores al modelo (biológicos, psicológicos y sociales) promoviendo esto, a su vez, el desarrollo de explicaciones híbridas y pluralistas.

Como argumenté, incluso hablar de dos factores continúa resultando insuficiente para dar cuenta de la complejidad de la producción y el mantenimiento de los delirios. Tanto el enfoque de uno como el de dos factores pretenden explicar solo la creencia delirante (o conjunto de creencias delirantes), enfatizando las características epistémicas del fenómeno y dejando de lado el contexto psicopatológico en la que esta emerge y la relación que mantiene con la vida psíquica global del sujeto, desatendiendo así el papel de los afectos en la génesis y el mantenimiento de los delirios. Frente a esto, afirmé que para comprender adecuadamente al delirio es necesario atender tanto al contexto general de afectividad negativa en el que emerge, como a la fenomenología de las alteraciones perceptuales-afectivas que suceden en lo que se ha denominado la atmósfera pre-delirante. Esto permitirá considerar el carácter procesual del fenómeno, entendiendo que la psicopatología del delirio abarca desde experiencias primarias agudas de la etapa pre-delirante, hasta el momento en el cual el delirio se torna crónico y las creencias delirantes se cristalizan e integran a la vida psicológica del sujeto. Asimismo, al apoyarse este modelo en un marco conceptual general de vulnerabilidad-estrés que considere además de los factores predisponentes y desencadenantes, el fenómeno puede ser abordado desde una perspectiva integral e interdisciplinaria que se valga de los aportes teóricos y empíricos de disciplinas tales como la psicología cognitiva, la psiquiatría, las neurociencias y la fenomenología.

Un punto de importancia a señalar es que, dada la evidencia disponible, no es posible determinar si todos los delirios involucrarían todos los tipos de factores mencionados (por ejemplo, no podemos aún determinar si todos los delirios estarían precedidos por una etapa pre-delirante). Esto hace referencia a la cuestión crucial de si todos los delirios son similares estructuralmente en relación a su etiología. Como ha señalado Berríos (2008), por lo que sabemos hasta el momento, aún no hay una clara respuesta empírica acerca de si el delirio es el mismo fenómeno en la

paranoia, esquizofrenia, en las manías o en los delirios orgánicos (*delirium tremens)* o si, eventualmente, existirá una meta-teoría que permitirá establecer claras distinciones. A fin de cuentas, los delirios pertenecientes a enfermedades diferentes podrían ser fenómenos diferentes. Si bien las respuestas brindadas a los diversos interrogantes que nos he planteado aún son incompletas, se espera que puedan servir como guía heurística. Serán eventualmente la investigación clínica y la producción de evidencia empírica las que determinarán la aplicabilidad de un modelo multifactorial como el propuesto a diferentes tipos de delirios. En este sentido, si bien el modelo presentado identifica varios factores que podrían formar y mantener los delirios, ha sido esquematizado en un nivel de generalidad tal que permita eventualmente realizar evaluaciones cuidadosas para identificar en los diversos casos individuales qué factores particulares se encuentran presentes operando (¿qué factores de vulnerabilidad están presentes?, ¿qué afectos se experimentan y de qué modo?, ¿qué tipo de explicaciones brinda el sujeto? etc.). En este sentido, será necesario profundizar en la investigación de los factores causales que operan en cada caso particular.

Referencias

American Psychiatric Association [APA] (2013). *Diagnostic and statistical manual of mental disorders: DSM-V-TR* (5th ed., text revision). Washington: American Psychiatric Association.

Anderson, B. (2009). Affective atmospheres. *Emotion, Space and Society*, 2, pp. 77-81.

Bayne, T. y Pacherie, E. (2004). Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. Philosophy, *Psychiatry y Psychology*, 11 (1), pp. 1-11.

Bell, D. (2003). *Paranoia.* UK: Icon.

Bentall, R. P. y Kaney, S. (1996). Abnormalities of self-representation and persecutory delusions: a test of a cognitive model of paranoia. *Psychological Medicine*, 26 (6), pp. 1231-1237.

Bentall, R. y Kinderman, P. (1998). Psychological processes and delusional beliefs: Implications for the treatment of paranoid states. In T. Wykes, N. Tarrier, y S. Lewis (Eds.), *Outcome and innovation in psychological treatment of schizophrenia*. US: Wiley.

Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N., and Kinderman, P. (2001). Persecutory delusions: A review and theoretical integration. *Clinical Psychology Review*, 21, pp. 1143-1192.

Bentall, R. (1995). Brains, biases, deficits and disorders. *British Journal of Psychiatry*, 167, pp. 153-55.

Berríos, G. (2008). Historia de los síntomas de los trastornos mentales. DF: Fondo de Cultura Económico.

Berríos, G. y Fuentenebro, F. (1995). Delirio: Historia, Clínica, Metateoría. Madrid: Trotta.

Bleuler, E. (1911/2011). *Dementia Praecox o el Grupo de las Esquizofrenias*. Buenos Aires: Polemos.

Bortolotti, L. (2014). The epistemic innocence of motivated delusions. *Consciousness y Cognition*, 33, pp. 490-499.

Broome, M., Day, F., Valli, I. (2012). Delusional ideation, manic symptomatology and working memory in a cohort at clinical high-risk for psychosis: a longitudinal study. *European Psychiatry*, 27, pp. 258-263.

Broome, M., Woolley, J., Tabraham, P. (2005). What causes the onset of psychosis? *Schizophrenia Research*, 79 (1), pp. 23-34.

Campbell, J. (2001). Rationality, meaning and the analysis of delusions. *Philosophy, Psychiatry y Psychology*, 8 (2), pp. 89-100.

Clark, D. (1999). Anxiety disorders: Why they persist and how to treat them. Behaviour Research and Therapy, 37, pp. 5-27.

Colheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *The Quarterly Journal of Experimental Psychology*, 60 (8), pp. 1041-1062.

Coltheart, M., Langdon R. y McKay, R. (2011). Delusional Belief. *Annu Rev Psychology*, 62, pp. 271-98, doi: 10.1146/annurev.psych.121208.131622.

Davies, M. y Coltheart, M. (2000). Introduction: Pathologies of Beliefs. In M. Coltheart and M. Davies (Eds.), *Pathologies of Beliefs*. UK: Blackwell.

Davies, M., Coltheart, M., Langdom, R. y Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry y Psychology*, 8, pp. 133-58.

Dickerson, F. (2000). Cognitive behavioural psychotherapy for schizophrenia: A review of recent empirical studies. *Schizophrenia Research*, 43 (2), pp. 71-90.

Dudley, R., John, A. y Young, A. (1997). Normal and abnormal reasoning in people with delusions. *British Journal of Clinical Psychology*, 36 (2), pp. 243-258.

Ellis, H. y Young, A. (1990). Accounting for delusional misidentifications. *British Journal of Psychiatry*, 157, pp. 239-248.

Ey, H. (1950*). Estudios sobre los delirios.* Madrid: Triacastela. Feinberg,

Firth, C. (1992). *The cognitive neuropsychology of schizophrenia*. Hove: IEA.

Fowler, D. (2000). Psychological formulation of early episodes of psychosis: a cognitive model. In M. Birchwood, D. Flowler y C. Jackson (Eds.), *Early intervention in psychosis: A guide to concepts, evidence and interventions*. Chichester: Wiley.

Freedman, B., y Chapman, L. J. (1973). Early subjective experience in schizophrenic episodes.

Journal of Abnormal Psychology, 82 (1), pp. 46-54.

Freeman, D., y Garety, P. A. (2002). Cognitive therapy for an individual with a long-standing: Incorporating emotional processes into a multi-factorial perspective on delusional beliefs. In: Morrison, A. (ed.), *A casebook of cognitive therapy for psychosis*. NY: Routledge.

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., y Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, 41, pp. 331-347.

Freeman, D., y Garety, P. A. (1999). Worry processes and dimensions of delusions: An exploratory investigation of a role for anxiety processes in the maintenance of delusional distress. *Behavioural and Cognitive Psychotherapy*¸ 27 (1), pp. 47-62.

Freeman, D., Garety, P. y Philips, M. (2000). An examination of hypervigilance for external threat in individuals with generalized anxiety disorder and individuals with persecutory delusions using visual scan paths. *Quarterly Journal of Experimental Psychology*, 53, pp. 549-567.

Freeman, D. (2016). Persecutory delusions: a cognitive perspective on understanding and treatment. The Lancet Psychiatry, 3, pp. 685-692.

Fuchs, T. (2013). The phenomenology of affectivity. In K. Fulford, M. Davies, R. Gipps, G., Graham, G., Sadler, J. Stanghellini, G. y Thorton, T. (Eds.), The Oxford Handbook of Philosophy and Psychiatry. Oxford: Oxford University Press.

—— (2005). Delusional mood and delusional perception. A phenomenological analysis. Psychopathology, 38, pp. 133-139.

Garety, P. y Freeman, D. (1999). Cognitive approaches to delusions: A critical review of theories and evidence. *British Journal of Clinical Psychology*, 38, pp. 113-54.

Garety, P. y Hemsley, D. (1997). *Delusions: Investigations into the Psychology of delusional Reasoning*. Oxford: Oxford University Press.

Garety, P., Hemsley, D. y Wessely, S. (1991). Reasoning in deluded schizophrenic and paranoid patients: Biases in performance on a probabilistic inference task. *Journal of Nervous and Mental Disease*, 179, pp. 194-201.

Garety, P., Kuipers, E., Fowler, D., Freeman, D. y Bebbington, P. E. (2001). A cognitive model of positive symptoms of psychosis. *Psychological Medicine*, 31 (2), pp. 189-95.

Gerrans, P. (1999). Delusional misidentification as subpersonal disintegration. *The Monist*, 82, pp. 590-608.

Gibbs, P. (2000). Thought Insertion and the Inseparability Thesis. *Philosophy, Psychiatry, y Psychology*, 7(3), pp. 195-202.

Goh, C. y Agius, M. (2010). The stress-vulnerability model how does stress impact on mental illness at the level of the brain and what are the consequences? *Psychiatry Danub*, 22 (2), pp. 198-202.

Jaspers, K. (1913/1963). *Psicopatología General*. Ed. Beta: Bs. As.

Kinderman, P. (1994). Attentional bias, persecutory delusions and the self concept. *British Journal of*

Medical Psychology, 67, pp. 53-66.

Klee, R. (2004). Why Some Delusions Are Necessarily Inexplicable Beliefs. *Philosophy, Psychiatry, y Psychology*, 11(1), pp. 25-34.

Kramer, I., Simons, C., Wigman, J., Collip, D., Jacobs, N., Derom, C., Thiery, E., van Os, J., Myinh-Germeys, I. y Wichers, M. (2013). Time-Lagged Moment-to-Moment Interplay Between Negative Affect and Paranoia: New Insights in the Affective Pathway to Psychosis. *Schizophrenia Bulletin*, 40 (2), pp. 278-286.

Langdom, R., Davies, M. y Coltheart, M. (2002). Understanding minds and understanding communicated meanings in schizophrenia. *Mind and Language*, 17, pp. 68-104.

Lincoln, T. M., Peter, N., Schäfer, M. y Moritz, S. (2010). From stress to paranoia: an experimental investigation of the moderating and mediating role of reasoning biases. *Psychological Medicine*, 40, pp. 169-171.

López, E. y Vázquez, C. (1988). *Modelo de vulnerabilidad y esquizofrenia*. Monografías de Psiquiatría, 1 (3), pp. 1-6.

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, 30, pp. 98-113.

—— (1988). Anomalous experience and delusional thinking: The logic of explanations. In *Delusional beliefs*, T. F. Oltmanns and B. A. Maher (Eds.). Chichester: John Wiley and Sons.

—— (1992). Delusions: Contemporary etiological hypotheses. *Psychiatric Annals*, 22, pp. 260-68.

—— (1999). Anomalous experience in everyday life: Its significance for psychopathology. The *Monist*, 2, pp. 547-70.

—— (2001). Delusions. In Patricia B. Sutker y Henry E. Adams (Eds.), *Comprehensive Handbook of Psychopathology*. Ciudad. Kluwer Academic.

Maiese, M. (2013). Thought insertion as a disownership symptom. *Phenomenology y the Cognitive Sciences*, 14 (4), pp. 911-927.

Marwaha, S. Broome, M., Bebbington, P., Kuipers, E. y Freeman, D. (2013). Mood Instability and Psychosis: Analyses of British National Survey Data. *Schizophrenia Bulletin*, 40 (2), pp. 269-277.

Mathews, A. (1990). Why worry? The cognitive function of anxiety. *Behaviour Research and Therapy*, 28, pp. 455-468.

McKay, R. y Cicopolotti, L. (2007). Attributional style in a case of Cotard delusion. *Consciousness and Cognition*, 16, pp. 349-359.

McKay, R., Langdon, R. y Colheart, M. (2009). “Sleights of mind: Delusions, and self deception”. In T. Bayne and J. Fernàndez (eds.), *Delusion and Self deception: Affective and Motivational Influences on Belief Formation*. Hove: Psychology Press.

Milosevic, I. y Radomsky, A. (2008). Safety behaviour does not necessarily interfere with exposure

therapy. *Behaviour Research and Therapy*, 46, pp. 1111-1118.

Mishara, A. L. (2010). Klaus Conrad (1905-1961): Delusional mood, psychosis and beginning Schizoprenia. *Schizophrenia Bulletin*, 36 (1), pp. 9-13.

Miyazono, K. y Bortolotti, L. (2015). Recent Work on the Nature and Development of Delusions. *Philosophy Compass*, *10*, 9, pp. 636-645.

Moritz, S. y Woodward, T. S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology*, 44, pp. 193-207.

—— (2006). A generalized bias against disconfirmatory evidence in schizophrenia. *Psychiatry Research*, 142, pp. 157-165.

Moritz, S. y Woodward, T. S. y Lambert, M. (2007). Under what circumstances do patients with schizophrenia jump to conclusions? A liberal acceptance account. *British Journal of Clinical Psychology*, 46, pp. 127-137.

Nelson, H. (2005). *Cognitive-behavioural Therapy with delusions and hallucinations: a practice manual*. NY: Nelson Thornes Ltd.

O'Driscoll, C., Laing, J. y Mason, O. (2014). Cognitive emotion regulation strategies, alexithymia and dissociation in schizophrenia, a review and meta-analysis. *Clinical Psychology Review*, 4, pp. 482-495.

Porcher, J. (2015). *Prospectos para uma teoria do delírio*. Tesis Doctoral. Universidade Federal do Río Grande do Soul. Porto Alegre, Brasil.

Read, J. Mosher, L. y Bentall, R. (2005). *Modelos de locura: aproximaciones psicológicas*. Madrid: Herder.

Saas, L. A. (1992). *Madness and modernism: Insanity in the light of modern art, literature, and thought*. Harvard: Harvard University Press.

Salwovsky, P. (1996). The cognitive approach to anxiety: Threath beliefs, safety-seeking behaviours, and the special case of health anxiety and obsessions. In P. Salkvoskis (Ed.), *Frontiers of cognitive therapy*. New York: Guilford Press.

Schreber, D. (1979). *Memorias de un enfermo nervioso*. Bs. As: Lohé.

Stone, T. y Young, A. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind y Language*, 12, pp. 327-64.

Strauss, G., Kappenman, E., Culbreth, A., Catalano, L., Lee, B. y Gold, J. (2013). Emotion Regulation Abnormalities in Schizophrenia: Cognitive Change Strategies Fail to Decrease the Neural Response to Unpleasant Stimuli. *Schizophrenia Bulletin*, 39 (4), 872-883.

Thornton, T. (2008) Why the idea of framework propositions cannot contribute to an understanding of delusions. *Phenomenology y the Cognitive Sciences*, 7 (2), pp. 159-175.

Torrelavega, R. (2016). Programa de Prevencion de la Psicosis [PPP] (Cantabria), España. Disponible online en: http://www.p3-info.es/view_article.asp?id=9ycat=3.

Trower, P. y Chadwick, P. (1995). Pathways to defense of the self: A theory of two types of paranoia. Clinical Psychology: *Science y Practice*, 2, pp. 263-278.

Turnbull, G. y Bebbington, P. (2001). Anxiety and the schizophrenic process: Clinical and epidemiological evidence. *Social y Psychiatric Epidemiology*, 36, pp. 235-243.

Veckenstedt, R., Randjbar, S., Vitzthum, F., Hottenrott, B., Woodward, T. S. y Moritz, S. (2011). Incorrigibility, jumping to conclusions, and decision threshold in schizophrenia. *Cognitive Neuropsychiatry*, 16, pp. 174-192.

Wahl, O. (1999). Mental health consumers' experience of stigma. *Schizophrenia Bulletin*, 25, pp. 467-478.

Wittgenstein, L. (1969). *On Certainty*. Oxford: Blackwell.

Woodward, T. S., Moritz, S. y Chen E. Y. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions: A study in an Asian sample with first episode schizophrenia spectrum disorders. *Schizophrenia Research*, 83, pp. 297-298.

Woodward, T. S., Moritz, S., Menon, M., Klinge, R. (2008). Belief inflexibility in schizophrenia. *Cognitive Neuropsychiatry*, 13, pp. 267-277.

Young, A. (1998). *Face and Mind*. Oxford: Oxford University Press.

—— (2000). Wondrous strange: The neuropsychology of abnormal beliefs. *Mind y Language*, 15 (1), pp. 47-73.

Zubin, J. y Spring, B. (1997). Vulnerability: a new view of schizophrenia. *Journal of Abnormal Psychology*, 86, pp. 103-123.

---

[1] Se trata de los modelos basados en el carácter explicativo *("explanationist models")* y los modelos de respaldo ("*endorsement models*") respectivamente. Para los modelos basados en el carácter explicativo, el delirio es adoptado como un intento de explicar una experiencia inusual, mientras que para los modelos del respaldo, la experiencia comprende el contenido mismo del delirio.

[2] Un delirio monotemático según el cual el sujeto cree que un familiar afectivamente cercano (p. ej.: cónyuge, padre o madre) ha sido reemplazado por un impostor.

[3] Cabe aclarar que, en última instancia, tanto para el enfoque "*bottom-up*" como para este, la génesis del delirio residiría en un mal funcionamiento orgánico. En este sentido, podría considerarse que ambos enfoques son de tipo "*bottom-up*". No obstante, en el contexto de la discusión la diferencia está planteada en términos de la relación entre experiencia y creencias. El enfoque de Campbell, es denominado por él mismo como "*top-down*" porque, apelando a la contraposición creencia/percepción, considera a la creencia como el fenómeno racional de "alto nivel" que incide

sobre la percepción. No obstante, cabe señalar que esto parece implicar no reconocer del carácter básico y primitivo de las creencias fundamentales que se alterarían, al menos en la medida en las que se las análoga a las “proposiciones marco” wittgensteineanas.

[4] No es necesario leer a Campbell (2001) como pensando que la relación entre el daño cerebral y la creencia es causal. Como señalan Bayne y Pacherie (2004), podría pensarse también esta relación metafísica en términos de superveniencia más que de causación.

[5] La razón es la siguiente: que el sujeto delirante sienta una respuesta afectiva inusual ante su cónyuge (la falta de familiaridad) no constituye, por sí mismo, una experiencia perceptual con un contenido anómalo (p. ej.: “Esta mujer que estoy viendo no es esa mujer que recuerdo como mi esposa”). Pero, solo una experiencia perceptual con un contenido como el recién mencionado (u otro similar) podría justificar el paso del sujeto a la creencia delirante.

[6] Es decir, la fase inicial del desarrollo del trastorno, donde aparecen los síntomas iniciales y el sujeto experimenta distintos tipos de cambios cognitivos, perceptivos y/o afectivos.

[7] Véase Goh y Agius (2010) para una revisión del desarrollo de este marco conceptual.

[8] Véase Garety, Hemsley y Wessely (1991) para una revisión de dichos sesgos.

AUTORAS Y AUTORES

Pablo López-Silva

Obtuvo una licenciatura en psicología, título profesional de psicólogo y un diploma en Apego y Cuidados Tempranos en la Escuela de Psicología de la Pontificia Universidad Católica de Valparaíso. Durante ese período, se dedicó a estudiar la psicogénesis y algunos problemas conceptuales fundamentales del *self*. Luego, obtuvo un master in Research y un doctorado en Filosofía por la Universidad de Manchester, Reino Unido. Desde ese momento comenzó a investigar en el campo de la psicopatología, la fenomenología y la filosofía de la mente desde un punto de vista analítico. Pablo López-Silva posee diversas publicaciones indexadas en inglés y español, ha sido investigador Fondecyt y profesor visitante en la Universidad Rikkyo en Japón, Institut Jean Nicod de París, en la Universidad de Warwick en Reino Unido y en la Universidad de Illinois-Chicago en los Estados Unidos.

Francisco Osorio

Es antropólogo y doctor en filosofía por la Universidad de Chile. Creador y director de la revista *Cinta de Moebio*, especializada en epistemología de las ciencias sociales. Fue *Fulbright Fellow* en The University of Pennsylvania, profesor visitante honorario de la Universidad de Manchester y profesor de pregrado y postgrado en Sheffield Hallam University. Ha realizado varias colaboraciones con académicos latinoamericanos, participando en cursos de posgrado, conferencias, congresos y escribiendo artículos y libros en conjunto con ellos. Creó el Foro de Editores Científicos de Chile que reúne a los editores de revistas científicas para colaboración y apoyo a la labor de comunicación científica.

Tim Bayne

Es uno de los filósofos de la mente y ciencias cognitivas más reconocidos a nivel mundial. Actualmente es *Professor of Philosophy* en la Universidad Monash, Australia. Previamente se desarrolló como académico e investigador en la Universidad de Oxford, Universidad Macquarie, Universidad de Western Ontario y Universidad de Manchester. Su actual investigación intenta explorar las formas en que la conciencia podría medirse y cómo este asunto impactaría a las actuales ciencias cognitivas.

Marcelo Díaz

Es académico y director del departamento de filosofía de la Universidad de Santiago de Chile. Desde sus estudios doctorales, su trabajo académico se ha centrado en la filosofía del lenguaje de Wittgenstein y la filosofía de la mente con especial énfasis en el estudio de la conciencia y el concepto de lo mental.

Otto Dörr

Ha realizado gran parte de su vida académica en la Universidad de Chile, especialmente en la Facultad de Medicina y en el Centro de Estudios de Fenomenología y Psiquiatría de la Universidad Diego Portales. Fue el primer médico en diagnosticar la bulimia nerviosa (1972). Sus publicaciones científicas son numerosas, así como su participación en congresos, seminarios, conferencias en diversas universidades del mundo, junto a otras actividades de docencia, investigación y extensión.

Roberto Ferrari

Biólogo y entomólogo. Académico de la Universidad de Boloña. Sus

intereses actuales están en la etología cognitiva de las mentes colectivas, la percepción de ambientes naturales, la experiencia animal, los derechos de los animales y el enfoque enactivo de Francisco Varela. Participa en el Centro de Estudios de Asia, donde realiza una comparación entre la investigación científica sobre la conciencia y las tradiciones experimentales de Oriente.

Alfredo Gaete

Es psicólogo y doctor en Filosofía por la Universidad de Manchester, Reino Unido. Su investigación filosófica se centra en el concepto de lo mental especialmente en la figura de Ryle. Actualmente, es académico de la Pontificia Universidad de Chile, campus Villarrica, donde realiza investigación en educación.

Rodrigo González

Realizo sus estudios de filosofía en la Universidad de Chile, de la cual actualmente es profesor del Departamento de Filosofía. Realizó su doctorado en la Universidad de Lovaina, Bélgica. Sus extensas publicaciones abarcan los temas de filosofía de la inteligencia artificial, filosofía de la mente, epistemología, filosofía del lenguaje, entre otras áreas.

Edmundo Kronmuller

Es psicólogo y doctor en Psicología por la Universidad de California-Riverside, Estados Unidos. Sus publicaciones y presentaciones en congresos son amplias y su investigación actual se centra en los mecanismos y condiciones cognitivas para la comprensión y coordinación lingüística.

Ricardo Pulido

Es psicólogo clínico por la Pontificia Universidad Católica de Chile. Psicoterapeuta relacional. Doctor en Psicología General y Clínica de la Universidad de Boloña. En la Universidad Alberto Hurtado ha sido director del Magíster de Estudios Sistémicos Relacionales de la Familia y de la Pareja. También director de la Unidad Mindfulness e instructor de meditación de la Asociación Spazio Interiore e Ambiente (A.S.I.A.).

María Ayelén Sánchez

Es doctora en filosofía por la Universidad Nacional del Sur (Argentina). El tópico general de su investigación filosófica es la atribución y la auto-atribución de intencionalidad, abordado desde la filosofía en diálogo con algunos de los resultados más relevantes del campo de la psicología cognitiva.

Alejandro Serani

Se graduó como médico cirujano por la Universidad de Chile y posteriormente realizó sus estudios de filosofía en Francia, en la Universidad de Toulouse. Posteriormente se especializó en Neurología Clínica. Es autor de gran cantidad de artículos y libros en neurociencia, bioética clínica y medicina, antropología médica y filosofía.

Emilia Vilatta

Es doctora en filosofía por la Universidad Nacional de Córdoba, Argentina, donde actualmente desarrolla su Carrera como investigadora post-doctoral. Su campo de trabajo es la filosofía de la mente y la filosofía de la psiquiatría con un especial énfasis en la naturaleza de los delirios.